Longitudinal Differential Item Functioning Detection Using Bifactor Models and the Wald Test

BY

Mian Wang

A dissertation submitted to the graduate degree program in the Department of Psychology and the Graduate Faculty of the University of Kansas in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Chairperson: Wei Wu

Co-Chairperson: Carol M. Woods

Ruth Ann Atchley

Pascal R. Deboeck

Jonathan Templin

Date defended: May 16th, 2016

The Dissertation Committee for Mian Wang

certifies that this is the approved version of the following dissertation:

Longitudinal Differential Item Functioning Detection Using Bifactor Models and the Wald Test

_____

Chairperson: Wei Wu

_____

Co-Chairperson: Carol M. Woods

Date approved: May 16th, 2016

Abstract

The use of longitudinal data for studying cross-time changes is built on the key assumption that properties (e.g., slopes and intercepts) of the repeatedly-used items remain unchanged over time. True changes in the latent variables are indistinguishable from item-level changes when items exhibit differential item functioning (DIF) across time points. To date, no research has extended the modified Wald test for longitudinal DIF detection. The current Monte Carlo simulation study proposes and evaluates a new approach, which pairs the versatile bifactor model with the modified Wald test, for detecting longitudinal DIF. Power and Type I error associated with DIF tests under the new approach are reported for conditions with varied proportions of known anchors and different types of standard error estimation procedure. The new approach is also compared to DIF methods based on the misspecified unidimensional model which assumes independence in the factors and items. An applied example is provided, along with the flexMIRT script and the R code used respectively for model calibration and DIF analysis. Limitations of the current study and future research directions are discussed.

Acknowledgements

First, I would like to thank my former adviser, Dr. Carol Woods, for her continued support and guidance throughout the years of my graduate career at the University of Kansas. I would have not completed my dissertation without her thorough feedback and insightful advice.

I also would like to thank Dr. Wei Wu for chairing my dissertation committee, Dr. Ruth Ann Atchley and Dr. Pascal Deboeck for serving as in-department committee members, Dr. Jonathan Templin for being an outside member, and Dr. David Johnson for serving on my proposal committee. Each of these professors has made contributions to the successful completion of my dissertation, and I am truly grateful for their valuable time and support.

A special thank you goes to Dr. Li Cai at the University of California, Los Angeles for providing his expert opinions on the simulation design. His work in the field of item response theory has inspired me with many research ideas including this dissertation.

Finally, I am deeply indebted to my mother, sister, and significant other for their love and encouragement. If it were not for them, I would have not persisted in accomplishing my doctoral degree.

Table of Contents

INTRODUCTION **1**

MOTIVATION 1

THE WALD TEST FOR DIF DETECTION 3

The unidimensional graded model. *3*

Basic Wald procedures for between-group DIF detection *4*

Modified Wald for simultaneous comparisons of multiple groups *6*

A MULTIDIMENSIONAL APPROACH FOR MODELING LONGITUDINAL DATA 9

The bifactor structure. *9*

Maximum marginal likelihood (MML) via expectation-maximization (EM). *10*

STANDARD ERROR APPROXIMATION PROCEDURES 17

The Fisher information matrix. *18*

Supplemented expectation maximization (SEM). *19*

Empirical cross-product approximation (XPD) *20*

THE CURRENT STUDY 21

The proposed approach. *21*

Comparisons of estimation methods. *21*

METHOD **22**

STUDY CONDITIONS 22

Manipulated variable. *22*

Fixed variables. *23*

Random variables. *25*

Miscellaneous. *26*

DATA GENERATION 26

Primary factors. *26*

Specific factors *27*

Item slopes (primary and specific). *27*

Item intercepts. *28*

DIF effects. *29*

PROCEDURE 29

EVALUATED OUTCOMES 30

Bias. *31*

Statistical power and Type I error. *36*

Computation time. *36*

RESULTS 36

CATEGORY COLLAPSE 36

CONVERGENCE OF MODEL CALIBRATION AND ERROR ESTIMATION 37

BIAS 39

Latent mean estimation. *39*

Latent SD estimation. *40*

Latent correlation estimation. *41*

Primary slope and intercept estimation. *41*

Specific slope estimation. *46*

STATISTICAL POWER AND TYPE I ERROR 46

Omnibus DIF test. *47*

Pairwise DIF tests. *48*

COMPUTATION TIME 49

EMPIRICAL EXAMPLE 50

DATA PROPERTIES 50

ANALYSIS PROCEDURES 51

DIF TEST RESULTS 52

DISCUSSION 52

FINDINGS OF THE CURRENT SIMULATION 53

SUGGESTIONS FOR APPLIED RESEARCH 54

LIMITATIONS AND FUTURE DIRECTIONS 56

APPENDIX A 58

APPENDIX B 83

APPENDIX C 86

APPENDIX D 88

BIBLIOGRAPHY 91

List of Figures

Figure 1. Path diagram of a bifactor structure for modeling longitudinal data collected from three time points. 9

Figure 2. Densities of Latent Mean Estimates at Time-2. 58

Figure 3. Densities of Latent Mean Estimates at Time-3. 59

Figure 4. RMSEs of Latent Mean Estimation.....60

Figure 5. MAPEs of Latent Mean Estimation. 61

Figure 6. Densities of Latent SD Estimates at Time-2. 62

Figure 7. Densities of Latent SD Estimates at Time-3. 63

Figure 8. RMSEs of Latent SD Estimation.....64

Figure 9. MAPEs of Latent SD Estimation. 65

Figure 10. Densities of Latent Correlation Estimates (IBA only). 66

Figure 11. RMSEs and MAPEs of Latent Correlation Estimates (IBA only).....67

Figure 12. RMSEs and MAPEs of Primary Parameter Estimates for Anchors (One-Anchor).....68

Figure 13. RMSEs and MAPEs of Primary Parameter Estimates for Anchors (Two-Anchor). 69

Figure 14. RMSEs and MAPEs of Primary Parameter Estimates for Time-1 Non-Anchors (One-Anchor). 70

Figure 15. RMSEs and MAPEs of Primary Parameter Estimates for Time-1 Non-Anchors (Two-Anchor). 71

Figure 16. RMSEs and MAPEs of Primary Parameter Estimates for Time-2 Non-Anchors (One-Anchor). 72

Figure 17. RMSEs and MAPEs of Primary Parameter Estimates for Time-2 Non-Anchors (Two-Anchor). 73

Figure 18. Absolute Percentage Deviations of All IBA-Estimated Primary Parameters for Time-2 Non-Anchors. ...74

Figure 19. RMSEs and MAPEs of Primary Parameter Estimates for Time-3 Non-Anchors (One-Anchor). 75

Figure 20. RMSEs and MAPEs of Primary Parameter Estimates for Time-3 Non-Anchors (Two-Anchor). 76

Figure 21. RMSEs of Specific Slope Estimates (for IBA Only).....77

Figure 22. MAPEs of Specific Slope Estimates (for IBA Only). 78

Figure 23. Statistical Power and Type I Error of Omnibus DIF Detection.....79

Figure 24. Statistical Power and Type I Error of DIF Detection Comparing Time-1 and Time-2. 80

Figure 25. Statistical Power and Type I Error of DIF Detection Comparing Time-1 and Time-3. 81

Figure 26. Computation Time Required for Model Estimation and Standard Error Calculation. 82

List of Tables

Table 1. Average Type I Errors of Omnibus DIF Tests.....83

Table 2. Average Type I Errors of Pairwise DIF Tests between Time-1 and Time-2. 83

Table 3. Average Type I Errors of Pairwise DIF Tests between Time-1 and Time-3. 83

Table 4. Primary Item Parameter Estimates for the 10-Item MFQ Subscale (Two Waves). 84

Table 5. IBA-XPD Results for Testing the 10-Item MFQ Subscale across Two Waves.....84

Table 6. IBA-SEM Results for Testing the 10-Item MFQ Subscale across Two Waves. 85

Introduction

Longitudinal studies are common nowadays within social scientific research areas. Researchers interested in assessing changes of constructs over time can test their hypotheses based on data collected from the same individuals on the same set of items across multiple time points. When modern latent variable modeling procedures are used, longitudinal changes in the unobservable latent constructs (indicated by observed items) are examined. However, a key underlying assumption for conducting latent-variable-based longitudinal research is that item properties (e.g., slopes and intercepts) do not vary across time. Violation of this assumption renders changes in the latent constructs indistinguishable from changes in the item properties. For example, improved scores on a non-verbal cognitive assessment test spanning childhood and adolescence might be partially due to improved reading skills (which would lead to better comprehension of questions and thus shifted item intercepts and/or slopes) rather than true changes of the non-verbal cognitive ability. Therefore, ensuring the time-invariant properties of items is a requisite step for conducting longitudinal studies.

Motivation

Under the item response theory (IRT) framework, measurement invariance is achieved by detecting and eliminating DIF effects (Embretson & Reise, 2000). Traditionally, DIF research has mainly focused on the analysis of data collected from two or more independent groups. For between-group DIF analysis, there is no need to consider the dependency of corresponding factors and items across groups, because all individuals and items are (assumed to be) independent according to the fundamental assumptions made by unidimensional IRT models. In contrast, due to the repeated measures nature of longitudinal data, the correlations between reoccurrences of the same factor (or the same item) should be taken into consideration before

DIF analysis is conducted. Once the cross-time dependencies are accounted for by appropriate multidimensional IRT models, the actual DIF analysis procedure is theoretically the same as in the multiple-group case – it detects differences between two or more sets of parameters associated with the same items. Thus, a method that works for between-group DIF analysis (e.g., the modified Wald test; Kim, Cohen, & Park, 1995; Langer, 2008; Lord, 1977; 1980; Wald, 1943; Woods, Cai, & Wang, 2013) can be adapted for longitudinal DIF detection.

To date, no research in the DIF literature has extended the modified Wald test for the detection of longitudinal DIF. Therefore, the focus of the current Monte Carlo simulation is on the formulation of a new strategy for detecting longitudinal DIF. There are three components of the proposed strategy: 1) Item parameters are estimated using the item bifactor analysis (IBA) models (Cai, 2010; Cai, Yang & Hansen, 2011; Gibbons & Hedeker, 1992; Hill, 2006) with designated time-invariant anchors, where cross-time dependency is modeled with regard to both the repeatedly measured items and the reoccurring latent factor; 2) Standard errors of the item parameters are approximated using either the SEM or the XPD procedure (Cai, 2008; Paek & Cai, 2014; Meng & Rubin, 1991); and 3) Implement the modified version of the Wald test (Cai, 2015; Kim et al., 1995; Langer, 2008; Lord, 1977; 1980; Wald, 1943) for DIF analysis of multiple time points, using estimated item parameters and standard errors obtained from the previous two steps.

In the following sections, we will first review the modified Wald test and its applications in between-group DIF analysis, and then provide details about the new strategy which extends the modified Wald test for DIF detection in longitudinal contexts.

The Wald Test for DIF Detection

The unidimensional graded model. Let $u_{ij} \in \{0, 1, \dots, V-1\}$ be the item response of person *i* on item *j* with V number of strictly ordered categories (e.g., a Likert-type scale; Likert, 1932), the unidimensional graded response model (GRM; Samejima, 1969) states that the item response functions (IRFs) for cumulative response probabilities given the person's level on the latent variable θ are

$$P(u_{ij} \geq 1|\theta_i) = \frac{1}{1+exp\{-(d_{j,1}+a_j\theta_i)\}}, \quad (1)$$

$$\vdots$$

$$P(u_{ij} \geq V-1|\theta_i) = \frac{1}{1+exp\{-(d_{j,V-1}+a_j\theta_i)\}}, \quad (2)$$

where $d_{j,1}, \dots, d_{j,V-1}$ are a set of intercepts for item *j*, and a_j is the item slope on the latent factor. In the unidimensional case, the slope-intercept and the traditional slope-threshold forms are interchangeable as the threshold parameter $b_{j,*} = -d_{j,*}/a_j$. The slope-intercept form is adopted to facilitate later discussions of multidimensional IRT models.

Consequently, the probability of responding in a given category is calculated as the difference between two cumulative probabilities. For example, the probability of choosing the third category $P(u_{ij} = 3|\theta_i)$ is simply the difference between two IRFs

$$P(u_{ij} \geq 3|\theta_i) - P(u_{ij} \geq 4|\theta_i) = \frac{1}{1+exp\{-(d_{j,3}+a_j\theta_i)\}} - \frac{1}{1+exp\{-(d_{j,4}+a_j\theta_i)\}}. \quad (3)$$

And the two boundaries of choosing the first and the V^{th} categories are respectively defined as $P(u_{ij} = 0|\theta_i) = 1 - P(u_{ij} \geq 1|\theta_i)$ and $P(u_{ij} = V-1|\theta_i) = P(u_{ij} \geq V-1|\theta_i)$, because $P(u_{ij} \geq 0|\theta_i) = 1$ and $P(u_{ij} \geq V|\theta_i) = 0$ by definitions of probability.

Under the special case where a GRM item has only two categories, its model can be reduced to a 2-parameter logistic (2PL) model, where the probability of person *i* responding

correctly on item j (denoted by $u_{ij} = 1$) given his/her level on the latent construct θ follows the IRF

$$P(u_{ij} = 1|\theta_i) = \frac{1}{1+exp\{-(d_j+a_j\theta_i)\}}, \quad (4)$$

where d_j is the intercept, and a_j is the slope for item j.

Basic Wald procedures for between-group DIF detection. Originally introduced by Lord (1977; 1980) for DIF detection, Lord's Wald test compares item parameters between two groups using Wald's (1943) χ^2 test statistic. The test in general fits a model in which group-invariant items (with their parameters fixed equal) are used for linking the two groups on the same latent metric, while parameters of studied items are freely estimated. After fitting appropriate IRT models to the items, two sets of parameters are obtained and the studied items are then compared for DIF detection purposes.

The Wald χ^2 statistic for the joint differences in item parameters between a reference (denoted by R in subscript) group and a focal group (denoted by F in subscript) is calculated as

$$\chi_j^2 = \left(\mathbf{v}_{j_R} - \mathbf{v}_{j_F}\right)^T \left(\mathbf{\Sigma}_{j_R} + \mathbf{\Sigma}_{j_F}\right)^{-1} \left(\mathbf{v}_{j_R} - \mathbf{v}_{j_F}\right), \quad (5)$$

in which $\left(\mathbf{v}_{j_R} - \mathbf{v}_{j_F}\right)^T$ holds item parameter differences (e.g., it holds $\left[\hat{a}_{j_R} - \hat{a}_{j_F}, \hat{d}_{j_R} - \hat{d}_{j_F}\right]$ for a 2PL item). The matrix $\left(\mathbf{\Sigma}_{j_R} + \mathbf{\Sigma}_{j_F}\right)$ is the sum of the asymptotic standard error covariance matrices associated with the item parameter estimates for both groups. For a 2PL item, the matrix is expressed as

$$\mathbf{\Sigma}_{j_R} + \mathbf{\Sigma}_{j_F} = \begin{bmatrix} var\left(\hat{a}_{j_R}\right) & cov\left(\hat{a}_{j_R}, \hat{d}_{j_R}\right) \\ cov\left(\hat{d}_{j_R}, \hat{a}_{j_R}\right) & var\left(\hat{d}_{j_R}\right) \end{bmatrix} + \begin{bmatrix} var\left(\hat{a}_{j_F}\right) & cov\left(\hat{a}_{j_F}, \hat{d}_{j_F}\right) \\ cov\left(\hat{d}_{j_F}, \hat{a}_{j_F}\right) & var\left(\hat{d}_{j_F}\right) \end{bmatrix}, \quad (6)$$

where $var\left(\hat{a}_{j_R}\right)$ and $var\left(\hat{a}_{j_F}\right)$ are the variances (i.e., squared standard error) of the estimated discrimination parameters for the focal and reference groups respectively, $var\left(\hat{d}_{j_R}\right)$ and

$var\left(\hat{d}_{j_F}\right)$ are the variances of the intercept parameters, and the off diagonal elements are covariances between the slope and intercept parameters for item *j*. For models with more than two parameters, the $\left(\mathbf{v}_{j_R} - \mathbf{v}_{j_F}\right)$ vector expands to include between-group differences in all parameters for item *j*, and the dimension of the error covariance matrix for each group also expands accordingly to accommodate the additional variances and covariances. For instance, an item that follows the GRM with five response categories will have a 5×1 vector for the parameter differences, and a 5×5 covariance matrix for each set of item parameters.

The degrees of freedom (*df*) for the aforementioned χ_j^2 equals the number of parameters being compared for the studied item (e.g., 2 for a 2PL item). An item is flagged as having a significant DIF effect when the test statistic is significant (e.g., $\chi_j^2(df = 2) > 5.991$ at $\alpha = .05$).

For the unconditional test of DIF in the slope parameter a_j, the Wald test statistic is calculated as

$$Z_{a_j}^2 = \frac{\left(\hat{a}_{j_R} - \hat{a}_{j_F}\right)^2}{\sigma^2_{\hat{a}_{j_R} - \hat{a}_{j_F}}}, \tag{7}$$

where $\sigma^2_{\hat{a}_{j_R} - \hat{a}_{j_F}}$ is the variance of the between-group differences in discrimination parameters, and $Z_{a_j}^2$ is chi-square distributed on 1 *df* for large samples. Significance of the $Z_{a_j}^2$ statistic indicates a significant nonuniform DIF effect of item *j* (i.e., item *j* have IRFs with different slopes for the two groups).

Conditioning on the equality of the a_j parameter between groups, the intercept parameter d_j can be test by calculating the difference between the overall Wald statistic and $Z_{a_j}^2$ with

$$Z_{d_j|a_j}^2 = \chi_j^2 - Z_{a_j}^2, \tag{8}$$

where $Z^2_{d_j|a_j}$ is also chi-square distributed on 1 *df* in a 2PL model for large samples. Significance of a $Z^2_{d_j|a_j}$ statistic indicates a significant DIF effect on the item intercept, which is equivalent to the test of the difficulty parameter b_j under unidimensional IRT models (i.e., detection of the uniform DIF effect). Note that, for a GRM with more than two categories, the test statistic $Z^2_{\{d_{j,1},\ldots,d_{j,v-1}\}|a_j} = \chi^2_j - Z^2_{a_j}$ detects the overall difference in the intercepts $\{d_{j,1}, \ldots, d_{j,v-1}\}$ conditioning on the equality of the slope a_j, and the *df* equals the total number of intercepts.

Modified Wald for simultaneous comparisons of multiple groups. With the limitation of comparing only two groups at a time, the original version of the Wald test was not an ideal solution for situations where multiple groups need to be tested for DIF. One can certainly run pairwise comparisons separately, but such practice results in undue work and concerns about inflated Type I error rates. Therefore, a modified version of the Wald test was introduced by Kim et al. (1995) that has the capability to compare multiple sets of item parameter estimates simultaneously with the inclusion of a contrast coefficient matrix (see also Langer, 2008; Woods et al., 2013).

The multiple-group DIF statistic that tests for the homogeneity of multiple sets of item parameters for item *j* is given by

$$Q_j = \left(\mathbf{C}\mathbf{v}_j\right)^T \left(\mathbf{C}\boldsymbol{\Sigma}_j\mathbf{C}^T\right)^{-1} \left(\mathbf{C}\mathbf{v}_j\right) \sim \chi^2_p \,, \tag{9}$$

where $\mathbf{v}_j$ is a vector of item parameter estimates for all groups, $\boldsymbol{\Sigma}_j$ is the block-diagonal nonsingular dispersion matrix of $\mathbf{v}_j$ (i.e., error covariance matrices of individual groups are placed along the diagonal with all other elements equal zero), and **C** is a matrix of linear contrasts which determines the pattern of comparisons being conducted. The test statistic Q_j is

chi-square distributed on *df* = *p*, where *p* is the rank of **C** (indicated by the total number of pivots in the contrast matrix; Dobson & Barnett, 2008).

Specifically, for a comparison of $K \in \{1, 2, \ldots, \mathrm{k}\}$ groups on a 2PL item,

$$\mathbf{v}_j = \left(\hat{a}_{j_1}\ \hat{d}_{j_1}\ \cdots\ \hat{a}_{j_k}\ \hat{d}_{j_k}\right)^T, \tag{10}$$

and

$$\boldsymbol{\Sigma}_j = \begin{bmatrix} var(\hat{a}_{j_1}) & cov(\hat{a}_{j_1}, \hat{d}_{j_1}) & \cdots & 0 & 0 \\ cov(\hat{d}_{j_1}, \hat{a}_{j_1}) & var(\hat{d}_{j_1}) & & 0 & 0 \\ \vdots & & \ddots & \vdots & \\ 0 & 0 & \cdots & var(\hat{a}_{j_k}) & cov(\hat{a}_{j_k}, \hat{d}_{j_k}) \\ 0 & 0 & & cov(\hat{d}_{j_k}, \hat{a}_{j_k}) & var(\hat{d}_{j_k}) \end{bmatrix}. \tag{11}$$

And the contrast matrix, prespecified by a researcher, can take different forms depending on how the parameters are compared across groups. Typically a reference group is chosen and all other groups are compared to the reference group. Therefore, the vector $\mathbf{C}\mathbf{v}_j$ contains between-group parameter differences compared in pairs. In a DIF study that compares parameters of a 2PL item between three groups with $\mathbf{v}_j = \left(\hat{a}_{j_1}\ \hat{d}_{j_1}\ \hat{a}_{j_2}\ \hat{d}_{j_2}\ \hat{a}_{j_3}\ \hat{d}_{j_3}\right)^{\mathrm{T}}$, for example, the contrast matrix in which all other groups are compared to the first group (i.e., the reference group) is constructed as

$$\mathbf{C} = \begin{bmatrix} 1 & 0 & -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 & 0 & -1 \end{bmatrix}, \tag{12}$$

and therefore

$$\mathbf{C}\mathbf{v}_j = \begin{bmatrix} \hat{a}_{j_1} - \hat{a}_{j_2} \\ \hat{d}_{j_1} - \hat{d}_{j_2} \\ \hat{a}_{j_1} - \hat{a}_{j_3} \\ \hat{d}_{j_1} - \hat{d}_{j_3} \end{bmatrix}. \tag{13}$$

Under situations where planned pairwise comparisons (e.g., between the first two groups) are conducted for the above example, with $\mathbf{v}_j = \left(\hat{a}_{j_1}\ \hat{d}_{j_1}\ \hat{a}_{j_2}\ \hat{d}_{j_2}\ \right)^{\mathrm{T}}$ and $\mathbf{C} = \begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{bmatrix}$, we have

$$\mathbf{C}\mathbf{v}_j = \begin{bmatrix} \hat{a}_{j_1} - \hat{a}_{j_2} \\ \hat{d}_{j_1} - \hat{d}_{j_2} \end{bmatrix}, \tag{14}$$

and

$$\mathbf{C}\boldsymbol{\Sigma}_j\mathbf{C}^T = \begin{bmatrix} var\left(\hat{a}_{j_1}\right) + var\left(\hat{a}_{j_2}\right) & cov\left(\hat{a}_{j_1}, \hat{d}_{j_1}\right) + cov\left(\hat{a}_{j_2}, \hat{d}_{j_2}\right) \\ cov\left(\hat{d}_{j_1}, \hat{a}_{j_1}\right) + cov\left(\hat{d}_{j_2}, \hat{a}_{j_2}\right) & var\left(\hat{d}_{j_1}\right) + var\left(\hat{d}_{j_2}\right) \end{bmatrix}. \tag{15}$$

Using the equation $Q_j = \left(\mathbf{C}\mathbf{v}_j\right)^{\mathrm{T}}\left(\mathbf{C}\boldsymbol{\Sigma}_j\mathbf{C}^{\mathrm{T}}\right)^{-1}\left(\mathbf{C}\mathbf{v}_j\right)$, the pairwise procedure yields a test statistic that is identical to Lord's original implementation of Wald test comparing two groups (with *df* = 2), shown in equation (5).

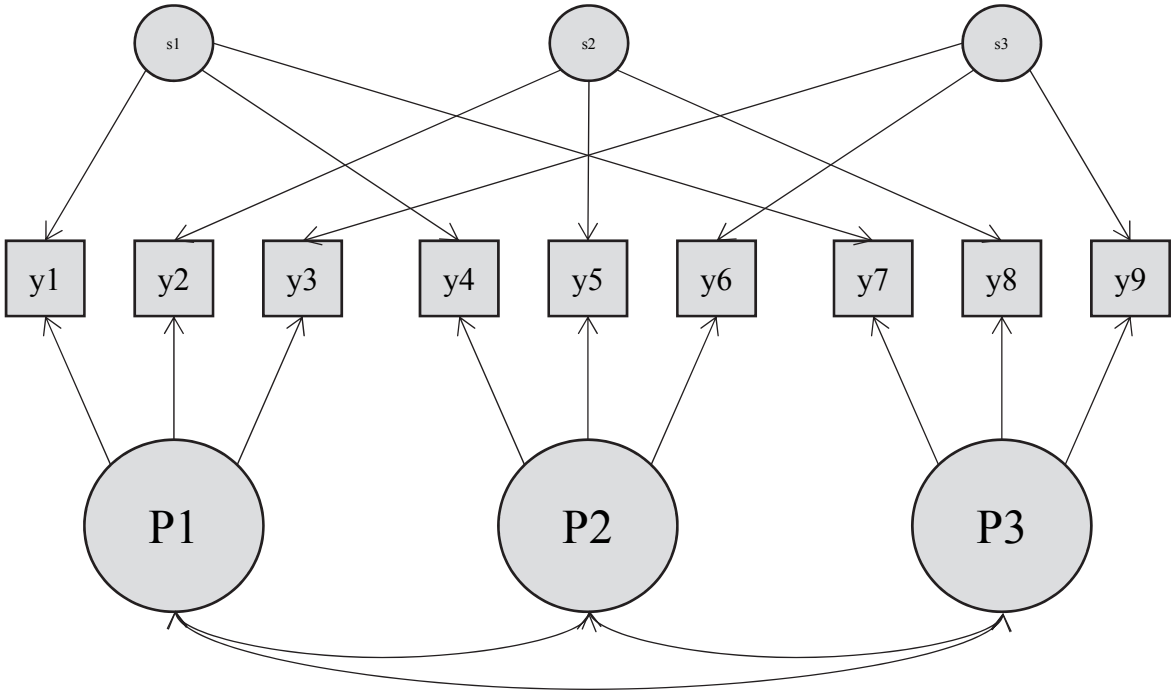
With some further improvements made for the linking and estimation procedures (Cai, 2015; Kim & Cohen, 2002; Kolen & Brennan, 2014; Langer, 2008), the modified version of the Wald test was shown to have adequate power and well-controlled Type I error for detecting DIF effect between two and three groups (Woods et al., 2013).

Even though the modified Wald test (along with the improvements) was designed for comparing item parameters of multiple independent groups, there is nothing preventing its application to DIF analysis in longitudinal contexts. As long as the item parameters are estimated using appropriate multidimensional IRT models (e.g., the IBA models; Cai, 2010; Cai et al., 2011; Gibbons & Hedeker, 1992; Hill, 2006) which relax the independence restrictions made by unidimensional IRT models, we should be able to recover accurate parameter estimates and standard errors for subsequent DIF analysis using the Wald test.

A Multidimensional Approach for Modeling Longitudinal Data

The bifactor structure. As mentioned earlier, due to the repeated measures nature of longitudinal data, the correlations between the reoccurrences of the same factor (and/or the same item) should be accounted for by a multidimensional IRT model. For example, a model for a three-item test taken by the same examinees at three time points can be depicted using a path diagram as shown below.

Figure 1. Path diagram of a bifactor structure for modeling longitudinal data collected from three time points.



In *Figure 1*, P1, P2 and P3 represent the construct of interest (i.e., the primary factor) modeled at three different time points. Variables y1, y4, and y7 represent the same item measured across three time points, and the same labeling scheme applies to the other items. The covariances (correlation if primary factors are standardized) between the repeated primary

factors are captured by the double-headed arrows. The cross-time residual covariations between the items are also accounted for by the introduction of a set of specific factors (e.g., y1, y4, and y7 are connected by s1 as a triplet).

The path diagram shown in *Figure 1* resembles a bifactor structure (Gibbons & Hedeker, 1992; Holzinger & Swineford, 1937) in which every item only loads on one primary factor and one specific factor. Thus, assuming two-category GRM items, we can write the IRF for the items shown in the diagram as

$$P\left(u_{ij}=1\middle|\theta_{i_P},\theta_{i_S}\right)=\frac{1}{1+exp\{-(d_j+a_{j_P}\theta_{i_P}+a_{j_S}\theta_{i_S})\}}, \tag{16}$$

where d_j is the item intercept, θ_{i_P} and θ_{i_S} are the latent scores of person *i* for the primary and specific factors respectively, and a_{j_P} and a_{j_S} are the slopes of item *j* for the corresponding factors.

Also, for the IBA extension of a *V*-category unidimensional GRM item (Cai et al., 2011), the cumulative probabilities of the ordered responses $u_{ij}\in\{0,1,\ldots,V-1\}$ is defined as

$$P\left(u_{ij}\geq 1\middle|\theta_{\mathrm{i_P}},\theta_{\mathrm{i_S}}\right)=\frac{1}{1+exp\{-\left(d_{j,1}+a_{j_P}\theta_{i_P}+a_{j_S}\theta_{i_S}\right)\}}, \tag{17}$$

$$\vdots$$

$$P\left(u_{ij}\geq V-1\middle|\theta_{i_P},\theta_{i_S}\right)=\frac{1}{1+exp\{-\left(d_{j,V-1}+a_{j_P}\theta_{i_P}+a_{j_S}\theta_{i_S}\right)\}}, \tag{18}$$

which closely resemble what we had in equations (1) and (2).

Maximum marginal likelihood (MML) via expectation-maximization (EM). Given our example of a three-item test carried out at three time points, the marginal likelihood for an individual with a response pattern **u** (with subscript *i* omitted for simplicity) on a 9-item test is

$$f_{\boldsymbol{\beta}}(\mathbf{u})=\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty}\left[\left(\prod_{j=1}^{9}f_{\boldsymbol{\beta}}\left(u_j\middle|\theta_{P1},\ldots,\theta_{s3}\right)\right)f(\theta_{P1},\ldots,\theta_{s3})\right]d\theta_{P1}\ldots d\theta_{s3}, \tag{19}$$

where **β** emphasizes the dependence of the likelihoods on a collection of unknown model parameters (e.g., item parameters, factor means, and factor variances). The notations in the form θ_{P1} denotes the distribution of the primary factor at the first time point (θ_{P2} and θ_{P3} are omitted for simplicity), and the θ_{s3} notation refers to the distribution of the third specific factor (θ_{s1} and θ_{s2} are omitted for simplicity).

Inside the brackets of equation (19), $f_{\boldsymbol{\beta}}(\mathrm{u}_j|\theta_{P1},\ldots,\theta_{s3})$ is the IRF (or the difference between two IRFs for a GRM item with more than two categories) for a particular response u_j on item *j* (see equations 16, 17, and 18), and thus

$$\prod_{j=1}^{9} f_{\boldsymbol{\beta}}(u_j|\theta_{P1},\ldots,\theta_{s3}) \tag{20}$$

is the product of all IRFs associated with a response pattern **u**. The distribution $f(\theta_{P1},\ldots,\theta_{s3})$ is the multivariate normal density of all six factors, including both primary and specific. By definitions of probability, the resulting product from inside the square brackets of equation (19) becomes the joint distribution $f_{\boldsymbol{\beta}}(\mathbf{u},\theta_{P1},\ldots,\theta_{s3})$. To obtain the marginal likelihood $f_{\boldsymbol{\beta}}(\mathbf{u})$, we need to integrate over the six dimensions as shown in equation (19), and it is typically approximated using summation over quadrature points in the form

$$f_{\boldsymbol{\beta}}(\mathbf{u}) \approx \sum_{q_{P1}=1}^{Q}\cdots\sum_{q_{s3}=1}^{Q}\left[\left(\prod_{j=1}^{9} f_{\boldsymbol{\beta}}(u_j|X_{q_{P1}},\ldots,X_{q_{s3}})\right)W_{q_{P1}}\ldots W_{q_{s3}}\right], \tag{21}$$

where *Q* is the number of quadrature nodes prespecified for the approximation procedure, each X_{q*} notation represents the corresponding abscissa value of a quadrature node on a given latent factor, and each W_{q*} notation represents the density weight at each node for the latent factor. The resulting quantity is the marginal likelihood (which can be considered as the volume under the surface in a six-dimensional space) associated with the response pattern **u** conditioning only on the unknown model parameters **β**. After item dependence on the primary factors and specific

factors being accounted for by the conditional likelihoods, the item responses are assumed to be independent.

Furthermore, let $\mathbf{U}$ be an $N \times J$ matrix where N is the sample size and J is the total number of items. To obtain the marginal likelihood of the observed data set $\mathbf{U}$, we need to take the product of the marginal likelihoods of all independent respondents as (with the previously omitted subscript i attached)

$$L(\boldsymbol{\beta}|\mathbf{U}) = \prod_{i=1}^{N} L(\boldsymbol{\beta}|\mathbf{u}_i) = \prod_{i=1}^{N} f_{\boldsymbol{\beta}}(\mathbf{u}_i)\,, \tag{22}$$

or expressed in the marginal log-likelihood form as

$$logL(\boldsymbol{\beta}|\mathbf{U}) = \sum_{i=1}^{N} logL(\boldsymbol{\beta}|\mathbf{u}_i) = \sum_{i=1}^{N} logf_{\boldsymbol{\beta}}(\mathbf{u}_i)\,, \tag{23}$$

where $\boldsymbol{\beta}$ refers to a vector of all unknown model parameters, and these parameters can be estimated by maximizing the log-likelihood function (see the EM algorithm section for details).

However, to calculate the nested sextuple summation in the marginal likelihood function using rectangular quadrature (as seen in equation 21), we need a total of Q^6 evaluations of the function inside the square brackets. The number of evaluations would grow exponentially with increasing numbers of quadrature points. For instance, more than 11 million evaluations would be required if we were to use 15 quadrature points per factor for the six-factor structure shown in *Figure 1*. Therefore, the estimation process soon becomes infeasible as the number of time points (which determines the number of primary factors) and the test length (which determines the number of specific factors) increase.

Dimension reduction. Motivated by Gibbons and Hedeker's (1992) work, Hill (2006) adapted the full-information IBA model for calibrating longitudinal data sets simulated for two time points, via a dimension reduction technique.

As shown in *Figure 1*, the specific factors are assumed to be uncorrelated with each other under the bifactor structure, and the primary factors are also uncorrelated with the specific

factors (though the primary factors themselves are correlated). Therefore we can utilize the property of conditional independence (Cai, 2010; Gibbons & Hedeker, 1992) to rewrite the product of all IRFs, with its original form shown in equation (20), associated with any given response pattern **u** as

$$\prod_{s\in\{s1,s2,s3\}}\prod_{j\in\mathbb{J}_s} f_{\boldsymbol{\beta}}\left(u_j|\theta_{P1},\dots,\theta_s\right) = f_{\boldsymbol{\beta}}(u_1,u_4,u_7|\theta_{P1},\theta_{P2},\theta_{P3},\theta_{s1})\times$$

$$f_{\boldsymbol{\beta}}(u_2,u_5,u_8|\theta_{P1},\theta_{P2},\theta_{P3},\theta_{s2})\times f_{\boldsymbol{\beta}}(u_3,u_6,u_9|\theta_{P1},\theta_{P2},\theta_{P3},\theta_{s3})\ , \qquad (24)$$

where *s* indexes each specific factor and $\mathbb{J}_s$ refers to the collection of items associated with a given specific factor.

Therefore, to obtain the marginal likelihood $f_{\boldsymbol{\beta}}(\mathbf{u})$, the three specific factors can be individually integrated out of their respective joint distribution first, and then followed by the integration over the three primary factors. This manipulation can be achieved using the equation

$$f_{\boldsymbol{\beta}}(\mathbf{u}) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\int_{+\infty}^{+\infty}[integrand]\ f(\theta_{P1},\theta_{P2},\theta_{P3})\,d\theta_{P1}d\theta_{P2}d\theta_{P3}\ , \qquad (25)$$

where the $[integrand]$ is the product of all marginal likelihoods for the specific factors in the form

$$[integrand] = f_{\boldsymbol{\beta}}(\mathbf{u}|\theta_{P1},\theta_{P2},\theta_{P3}) =$$

$$\left(\int_{-\infty}^{+\infty} f_{\boldsymbol{\beta}}(u_1,u_4,u_7|\theta_{P1},\theta_{P2},\theta_{P3},\theta_{s1})f(\theta_{s1})d\theta_{s1}\right)\times$$

$$\left(\int_{-\infty}^{+\infty} f_{\boldsymbol{\beta}}(u_2,u_5,u_8|\theta_{P1},\theta_{P2},\theta_{P3},\theta_{s2})f(\theta_{s2})d\theta_{s2}\right)\times$$

$$\left(\int_{-\infty}^{+\infty} f_{\boldsymbol{\beta}}(u_3,u_6,u_9|\theta_{P1},\theta_{P2},\theta_{P3},\theta_{s3})f(\theta_{s3})d\theta_{s3}\right)\ . \qquad (26)$$

As a result, when approximating the integral using rectangular quadrature, the $[integrand]$ only depends on the three primary factors because all specific factors inside the square brackets have been marginalized out. Then, the entire marginal likelihood, contingent upon the response pattern **u** of each subject, can be approximated as

$$f_{\boldsymbol{\beta}}(\mathbf{u}) \approx \sum_{q_{P1}=1}^{Q}\sum_{q_{P2}=1}^{Q}\sum_{q_{P3}=1}^{Q}\Big[\Big(\sum_{q_{s1}=1}^{Q} f_{\boldsymbol{\beta}}\big(u_1, u_4, u_7 \big| X_{q_{P1}}, X_{q_{P2}}, X_{q_{P3}}, X_{q_{s1}}\big) W_{q_{s1}}\Big) \times \Big(\sum_{q_{s2}=1}^{Q} f_{\boldsymbol{\beta}}\big(u_2, u_5, u_8 \big| X_{q_{P1}}, X_{q_{P2}}, X_{q_{P3}}, X_{q_{s2}}\big) W_{q_{s2}}\Big) \times \Big(\sum_{q_{s3}=1}^{Q} f_{\boldsymbol{\beta}}\big(u_3, u_6, u_9 \big| X_{q_{P1}}, X_{q_{P2}}, X_{q_{P3}}, X_{q_{s3}}\big) W_{q_{s3}}\Big)\Big] W_{q_{P1}} W_{q_{P2}} W_{q_{P3}} \,. \quad (27)$$

Thus the total number of required function evaluations has been reduced down to $s \times Q^{P+1} = 3 \times Q^4$, which saves considerable amount of computational time compared to other methods that do not perform dimension reduction (see Cai, 2010).

In addition, although the dimension reduction technique was originally built upon the IBA model's assumptions regarding the uncorrelated specific factors and dually-loaded items, Cai (2010) generalized this technique for the "bifactor-like" two-tier structure with some of the bifactor restrictions relaxed (e.g., items can load on more than two factors). Therefore, the dimension reduction technique is by no means limited to bifactor models, but also applicable for a wider range of multidimensional IRT models.

EM algorithm. The resulting marginal likelihood function from the previous section (i.e., equations 22 and 23 with or without dimension reduction) can be optimized with regard to **β** using the popular EM algorithm (Bock & Aitkin, 1981; Dempster, Laird, & Rubin, 1977) which consists of two iterative steps. At the expectation step (E-step), starting values for **β** are used, and the expected number of people in each response category are calculated at each quadrature point of each factor. These expected frequencies are then stored in an E-step table. In the maximization step (M-step), E-step data are used, as if they are complete data, to maximize the log-likelihood function. These two steps iterate until convergence on a given set of E-step and M-step criteria.

E-step. Recall that the marginal likelihood approximation of a response pattern **u** for person *i* is

$$L(\boldsymbol{\beta}|\mathbf{u}_i) = f_{\boldsymbol{\beta}}(\mathbf{u}_i) = \sum_{q_{P1}=1}^{Q}\sum_{q_{P2}=1}^{Q}\sum_{q_{P3}=1}^{Q}\{\prod_{s\in\{s1,s2,s3\}}[\sum_{q_s=1}^{Q}\prod_{j\in\mathbb{J}_s} f_{\boldsymbol{\beta}}(u_j|\theta_{P1},\theta_{P2},\theta_{P3},\theta_s)\, W_{q_s}]\}\, W_{q_{P1}} W_{q_{P2}} W_{q_{P3}}\,, \quad (28)$$

where $\prod_{j\in\mathbb{J}_s} f_{\boldsymbol{\beta}}(u_j|\theta_{P1},\theta_{P2},\theta_{P3},\theta_s)$ is the multidimensional ordinate of the product of all item characteristic surfaces for items associated with a given specific factor *s* at the corresponding quadrature points $(X_{q_{P1}}, X_{q_{P2}}, X_{q_{P3}}, X_{q_s})$ on the multidimensional abscissa (Cai, 2010). Also, let

$$E_i = \prod_{s\in\{s1,s2,s3\}}[\sum_{q_s=1}^{Q}\prod_{j\in\mathbb{J}_s} f_{\boldsymbol{\beta}}(u_j|\theta_{P1},\theta_{P2},\theta_{P3},\theta_s)\, W_{q_s}]\,, \quad (29)$$

and

$$E_{is} = \sum_{q_s=1}^{Q}\prod_{j\in\mathbb{J}_s} f_{\boldsymbol{\beta}}(u_j|\theta_{P1},\theta_{P2},\theta_{P3},\theta_s)\, W_{q_s} \quad (30)$$

be defined, we then have the following E-step entries computed:

1) The E-step expected frequencies for item *j* at quadrature points $(X_{q_{P1}}, X_{q_{P2}}, X_{q_{P3}}, X_{q_s})$, which depends on a specific factor *s*, is defined as

$$r_{ij}(X_{q_{P1}}, X_{q_{P2}}, X_{q_{P3}}, X_{q_s}) = \frac{E_i}{E_{is}} \times \frac{f_{\boldsymbol{\beta}}(u_{ij}|X_{q_{P1}}, X_{q_{P2}}, X_{q_{P3}}, X_{q_s})}{f_{\boldsymbol{\beta}}(\mathbf{u}_i)}\,. \quad (31)$$

   By summing over the total number of subjects respond in a given category (denoted by n_v), we have the expected frequencies for category *v* of item *j* as

$$r_{jv}(X_{q_{P1}}, X_{q_{P2}}, X_{q_{P3}}, X_{q_s}) = \sum_{i=1}^{n_v} r_{ij}(X_{q_{P1}}, X_{q_{P2}}, X_{q_{P3}}, X_{q_s})\,. \quad (32)$$

2) The E-step frequencies (of person *i*) for specific factor *s* can be defined as

$$r_{is}(X_{q_s}) = \frac{1}{f_{\boldsymbol{\beta}}(\mathbf{u}_i)}\sum_{q_{P1}=1}^{Q}\sum_{q_{P2}=1}^{Q}\sum_{q_{P3}=1}^{Q} f_{\boldsymbol{\beta}}(u_{ij}|X_{q_{P1}}, X_{q_{P2}}, X_{q_{P3}}, X_{q_s})\frac{E_i}{E_{is}} W_{q_{P1}} W_{q_{P2}} W_{q_{P3}}\,, \quad (33)$$

   and we obtain the frequencies at all quadrature points X_{q_s} for specific factor *s* by summing over all respondents

$$r_s(X_{q_s}) = \sum_{i=1}^{N} r_{is}(X_{q_s})\,. \quad (34)$$

3) The E-step frequencies (of person *i*) for primary factors can be defined as

$$r_{iP}(X_{q_{P1}}, X_{q_{P2}}, X_{q_{P3}}) = \frac{E_i}{f_{\boldsymbol{\beta}}(\mathbf{u}_i)}\,, \quad (35)$$

and we obtain the frequencies at all quadrature points $\left(X_{q_{P1}}, X_{q_{P2}}, X_{q_{P3}}\right)$ for the primary factors by summing over all respondents

$$r_P\left(X_{q_{P1}}, X_{q_{P2}}, X_{q_{P3}}\right) = \sum_{i=1}^{N} r_{iP}\left(X_{q_{P1}}, X_{q_{P2}}, X_{q_{P3}}\right). \quad (36)$$

M-step. Treating the E-step tables as the complete data, we can obtain model parameter estimates via the following likelihood equations (Note: $X_{q_{P2}}$ and its corresponding summation are omitted to conserve space):

1) The M-step log-likelihood function for item *j* is defined as

$$logL_j(\boldsymbol{\beta}) =$$

$$\sum_{v=1}^{V}\sum_{q_{P1}=1}^{Q}\cdots\sum_{q_{P3}=1}^{Q}\sum_{q_s=1}^{Q} r_{jv}\left(X_{q_{P1}}, \ldots, X_{q_{P3}}, X_{q_s}\right) logP_{jv}\left(X_{q_{P1}}, \ldots, X_{q_{P3}}, X_{q_s}, \boldsymbol{\beta}\right), \quad (37)$$

where V refers to the number of ordered categories (as previously defined) and $P_{jv}\left(X_{q_{P1}}, \ldots, X_{q_{P3}}, X_{q_s}, \boldsymbol{\beta}\right)$ is the IRF for each category of item *j* evaluated at quadrature points $\left(X_{q_{P1}}, \ldots, X_{q_{P3}}, X_{q_s}\right)$. By taking the derivative with respective to the unknown $\boldsymbol{\beta}$ on both sides of the equation, we have

$$\frac{\partial logL_j(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} =$$

$$\sum_{v=1}^{V}\sum_{q_{P1}=1}^{Q}\cdots\sum_{q_{P3}=1}^{Q}\sum_{q_s=1}^{Q}\left[\frac{r_{jv}\left(X_{q_{P1}},\ldots,X_{q_{P3}},X_{q_s}\right)}{P_{jv}\left(X_{q_{P1}},\ldots,X_{q_{P3}},X_{q_s},\boldsymbol{\beta}\right)} \times \frac{\partial P_{jv}\left(X_{q_{P1}},\ldots,X_{q_{P3}},X_{q_s},\boldsymbol{\beta}\right)}{\partial \boldsymbol{\beta}}\right]. \quad (38)$$

By setting the equation equal to zero, we can obtain the MML item parameter estimates via the iterative Newton-Raphson approximation method.

2) The M-step log-likelihood function for specific factor *s* is defined as

$$logL_s(\boldsymbol{\beta}) = \sum_{q_s=1}^{Q} r_s\left(X_{q_s}\right) logf\left(X_{q_s}|\boldsymbol{\beta}\right), \quad (39)$$

where $f\left(X_{q_s}|\boldsymbol{\beta}\right)$ is the univariate standard normal density (of specific factor *s*) evaluated at quadrature points X_{q_s}.

3) The M-step log-likelihood function for primary factors is defined as

$$logL_P(\boldsymbol{\beta}) = \sum_{q_{P1}=1}^{Q} \dots \sum_{q_{P3}=1}^{Q} r_P\left(X_{q_{P1}}, \dots, X_{q_{P3}}\right) logf\left(X_{q_{P1}}, \dots, X_{q_{P3}}|\boldsymbol{\beta}\right), \quad (40)$$

where $f\left(X_{q_{P1}}, \dots, X_{q_{P3}}|\boldsymbol{\beta}\right)$ is the multivariate normal density (of all three primary factors) evaluated at quadrature points $(X_{q_{P1}}, \dots, X_{q_{P3}})$.

Once the MML estimators of the unknown parameters are obtained for the current EM cycle, the provisional item parameters for the subsequent E-step are updated using the new estimates. The iterative process terminates when the change in the calculated marginal likelihoods between two E-step cycles is smaller than the convergence criterion (typically it is set at 0.001 or 0.0001), or the user-defined maximum allowed number of EM cycles is exhausted (see notes in the Method section).

Note that when a set of pre-identified anchors are available, we fix parameters of the anchor items equal across groups or time points. Therefore the **β** vector can be reduced to contain only the factor-level parameters for calibration purposes, if all items are known to be group-/time-invariant. Furthermore, in many situations where means and standard deviations (SDs) of some of the primary factors are unknown, we can standardize one of the primary factors (as the reference group or the reference time point), and estimating the means and SDs of the other factors in **β** as the unknown parameters (all latent factors are linked on the same metric using group-/time-invariant anchors).

Standard Error Approximation Procedures

As mentioned earlier, the Wald test requires a set of dispersion matrices which contain variances (i.e., squared standard errors) and covariances of all estimated item parameters. There is a variety of competing standard error estimation procedures, and each of them has its own advantages and disadvantages. The following sections will discuss three procedures that were evaluated by Paek and Cai (2014) under various conditions including the use of the traditional

bifactor models (i.e., items all load on a single primary factor and also on two or more uncorrelated specific factors). Two of the three procedures recommended by Paek and Cai (2014) are evaluated in the current simulation.

The Fisher information matrix. Ideally, the standard error covariance matrix is best computed using the Fisher information matrix approach which yields an unbiased item parameter covariance matrix (Paek & Cai, 2014; Tian, Cai, Thissen, & Xin, 2013).

Suppose there are a total of *K* distinct response patterns with $K = \prod_{j=1}^{J} v_j$ where v_j is the number of categories for item *j*, and *J* is the total number of items in the test. We can rewrite the marginal likelihood function from equation (22) in a format grouped by response patterns as

$$L(\boldsymbol{\beta}|\mathbf{U}) = \prod_{k=1}^{K} f_{\boldsymbol{\beta}}\left(\mathbf{u}_{i_k}\right)^{n_k}, \tag{41}$$

where *K* is the total number of distinct response patterns, $f_{\boldsymbol{\beta}}\left(\mathbf{u}_{i_k}\right)$ is the marginal likelihood associated with a specific pattern *k*, and n_k is the number of subjects who responded in pattern *k*.

By rearranging the collection of all $f_{\boldsymbol{\beta}}\left(\mathbf{u}_{i_k}\right)$ into a $K \times 1$ vector $\boldsymbol{\pi}(\boldsymbol{\beta})$, with $\boldsymbol{\beta}$ referring to a vector of unknown model parameters, we can calculate the Jacobian of the marginal likelihoods with regard to $\boldsymbol{\beta}$ as

$$\mathcal{J}(\boldsymbol{\beta}) = \frac{\partial \boldsymbol{\pi}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^{\mathrm{T}}}, \tag{42}$$

which contains first-order partial derivatives of log-likelihood functions (one for each response pattern) with respect to transposed unknown parameters (i.e., $\boldsymbol{\beta}^{\mathrm{T}}$). And the Fisher information matrix is defined as

$$\mathcal{F}(\boldsymbol{\beta}) = [\mathcal{J}(\boldsymbol{\beta})]^{\mathrm{T}}\{diag(\boldsymbol{\pi}(\boldsymbol{\beta}))\}^{-1}[\mathcal{J}(\boldsymbol{\beta})], \tag{43}$$

or it is sometimes calculated as -1 times the expected value of the Hessian matrix of $\boldsymbol{\beta}$ (contains second-order partial derivatives of the log-likelihood function)

$$\mathcal{F}(\boldsymbol{\beta}) = -E[H(\boldsymbol{\beta})] = -E\left[\frac{\partial^2 logL(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{\mathrm{T}}}\right]. \quad (44)$$

We can obtain the standard error covariance matrix as $\left[N * \mathcal{F}\left(\widehat{\boldsymbol{\beta}}\right)\right]^{-1}$ by evaluating $\mathcal{F}(\boldsymbol{\beta})$ at the model estimated parameters $\widehat{\boldsymbol{\beta}}$. The resulting covariance matrix, in which the diagonal elements are variances (squared standard errors) and the off-diagonal elements are covariances of the item parameter estimates, can be used in the aforementioned Wald test for DIF testing purposes.

One major drawback of the Fisher information matrix approach is the computation time it requires, especially with a long test (or short tests implemented at multiple time points) using multi-category GRM items. Hence, alternative standard error approximation procedures were sought by researchers in the past and some of the methods yielded desirable results based on simulation findings (Cai, 2008; Paek & Cai, 2014; Tian et al., 2013).

Supplemented expectation maximization (SEM). The SEM algorithm (Cai, 2008; Meng & Rubin, 1991) was developed based on the original idea of Smith published in the discussion section of Dempster et al. (1977). The large-sample covariance matrix from the EM-MML estimator is

$$\mathrm{COV}\left(\widehat{\boldsymbol{\beta}}|\mathbf{U}\right) = \mathcal{F}^{-1}\left(\widehat{\boldsymbol{\beta}}|\mathbf{U}\right) = \mathcal{F}^{-1}\left(\widehat{\boldsymbol{\beta}}\right) \cdot \left\{\mathbf{I}_d - \boldsymbol{\Delta}\left(\widehat{\boldsymbol{\beta}}\right)\right\}^{-1}, \quad (45)$$

where $\mathcal{F}^{-1}\left(\widehat{\boldsymbol{\beta}}|\mathbf{U}\right)$ is the inverted information matrix of the observed data **U**, $\mathcal{F}^{-1}\left(\widehat{\boldsymbol{\beta}}\right)$ is the inverted information matrix of the "pseudo-complete" data from an E-step, $\mathbf{I}_d$ is an identity matrix with dimension d equals the number of unknown parameters, and $\boldsymbol{\Delta}\left(\widehat{\boldsymbol{\beta}}\right)$ is the rate of convergence. The $\boldsymbol{\Delta}\left(\widehat{\boldsymbol{\beta}}\right)$ matrix is approximated iteratively using the entire output history from the original EM cycles. Tian et al. (2013) later introduced a more efficient version of SEM which uses only selective output from an optimal window of EM history. See Meng and Rubin

(1991) for more technical details regarding the SEM algorithm (see also Cai, 2008; Tian et al., 2013).

SEM (or its updated version) is one of the most efficient error covariance matrix estimation algorithms based on findings of previous research (Tian et al., 2013; Paek & Cai, 2014) under various simulation conditions. Its computation time is greatly reduced, in contrast to the Fisher information matrix approach, while it still maintains unbiased approximations of the standard errors. Moreover, research has shown that the modified Wald test, when used in conjunction with the SEM algorithm for detecting multiple-group DIF, maintained high power and well-controlled Type I error (Woods et al., 2013). However, due to the complexity of the multidimensional IRT models, the computational burden of the iterative SEM algorithm could get unwieldy. In a pilot run of the current simulation study, the SEM approximation in some replications took more than twice the time of the actual item parameter estimation process via EM-MML.

Empirical cross-product approximation (XPD). An even less computationally intensive procedure for standard error approximation is the XPD procedure where the information matrix is obtained as -1 times the expected value of the product of two matrices that contain first-order partial derivatives of the log-likelihood functions

$$\mathcal{F}(\boldsymbol{\beta}) = -E\left[\frac{\partial logL(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \cdot \frac{\partial logL(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^{\mathrm{T}}}\right]. \tag{46}$$

In the simulation study conducted by Paek and Cai (2014), the XPD method was found to be comparable to the Fisher information matrix and the SEM procedures, except under conditions where the sample-size-to-test-length ratio is small (e.g., 500 observations combined with a 40-item test). Nevertheless, XPD was the fastest in terms of computation time among the three procedures compared (i.e., Fisher, SEM, and XPD), and it always required less time than

the preceding item parameter estimation process. Therefore, XPD should be ideal for standard error computation under multidimensional IRT modeling, when the total sample size is substantially larger than the total number of items.

The Current Study

The proposed approach. Based on foregoing sections, we have the proposed longitudinal DIF detection method formulated in three steps: 1) Estimate model parameters via EM-MML using IBA models with anchor item parameters constrained equal across time; 2) Approximate standard errors using either SEM or XPD; and 3) Detect cross-time DIF effects using the modified Wald test designed to compare multiple sets of parameters simultaneously (with appropriate parameter estimates and standard errors obtained from the preceding steps). Successful implementation of this method on longitudinal data should yield desirable results in terms of power and Type I error associated with DIF detection.

Comparisons of estimation methods. In the present Monte Carlo simulation, parameter estimates and standard errors obtained from each of the four methods were subsequently analyzed using the modified Wald procedure: 1) the traditional multiple-group unidimensional GRM estimation with standard errors approximated using SEM ('MULTI-SEM' hereafter); 2) the multiple-group unidimensional GRM estimation with standard errors approximated using XPD ('MULTI-XPD' hereafter); 3) the IBA multidimensional GRM estimation with standard errors approximated using SEM ('IBA-SEM' hereafter); and 4) the IBA multidimensional GRM estimation with standard errors approximated using XPD ('IBA-XPD' hereafter).

The two MULTI-based methods were expected to perform much worse, in terms of model parameter and standard error estimation, than the two IBA-based methods due to the untenable independent-group and independent-item assumptions made by the MULTI-based

methods. The MULTI-based methods are included in the comparison in order to reveal potential advantages multidimensional methods have over unidimensional methods when data are truly longitudinal, although the two MULTI-based methods are not compared against each other given that they both are misspecified models. Moreover, the discrepancy between IBA-SEM and IBA-XPD procedures in terms of their power and Type I error associated with the subsequent DIF analysis, can be evaluated as references for weighing computation time against accuracy in applied research.

Method

Choices for the simulation conditions and data generating parameters described in the following sections were primarily informed by publications found in both the DIF literature (e.g., Meade & Wright, 2012; Wang & Yeh, 2003; Woods, 2009; Woods, 2011; Woods et al., 2013) and the measurement invariance literature (e.g., Bowers et al., 2010; Little, 2013; Pitts, West, & Tein, 1996; Short, 2014; Wu, Chen, & Tsai, 2009). Also, pilot simulations were conducted with 50 replications to acquire preliminary knowledge regarding the proposed study (e.g., choices of convergence criteria, required estimation time, and other relevant control variables).

Study Conditions

Manipulated variable. In order to control for estimation time in exchange for boosting the number of replications per condition, the only manipulated variable for the current study was the number of pre-identified anchor items. Either 10% or 20% of the total number of items was designated as time-invariant anchors in the estimation process. The chosen numbers/proportions for the anchor items were commonly seen in the previously mentioned DIF literature.

Fixed variables. To ensure that the simulation study completes in a timely manner and to control for unintended variability, the following aspects of the current study were held the same across all conditions:

i. Both the data-generating model and the IBA model followed the bifactor configuration (see *Figure 1* as an example). The two MULTI-based methods estimated the parameters by fitting unidimensional GRMs to the data, ignoring item and factor dependencies across time.

ii. All simulated items followed the IBA extension of the five-category GRM given its popularity in social scientific research. It is worth noting that, simulated data for the five-category ordinal items might not contain responses to all five categories due to various reasons such as a floor/ceiling effect. Instances of category collapsing that occurred during the current simulation were monitored, and proper procedures were carried out to accommodate items with collapsed categories (see Results section Category Collapse subsection for details).

iii. The number of time points simulated for the longitudinal design was always three (referred to as Time-1, Time-2 and Time-3 hereafter), as typically seen in the aforementioned measurement invariance literature.

iv. Simulated test length was fixed to be 10 (common for Likert-type psychological measures such as the Rosenberg Self-Esteem Scale; Rosenberg, 1965), and they were repeated at all three time points (i.e., no planned missing or other special manipulations involved).

v. The number of simulated respondents was fixed to be 1000 in order to maintain a relatively high sample-size-to-test-length ratio, and the number of observations per time point remained the same across time points (i.e., no dropout/attrition).

vi. The factor structure within each time point was always unidimensional (multidimensional factor structure within each time point is possible but it is computationally demanding under the full-information IBA models).

vii. Within each replication, items simulated as DIF-free (including time-invariant anchors and non-DIF studied items) at Time-1 remained DIF-free at subsequent time points.

viii. Within each replication, items simulated as having a DIF effect at Time-1 remained as DIF items at subsequent time points, and its true item parameters also varied between Time-2 and Time-3.

ix. Five hundred data sets were generated for each of the two anchoring conditions, and each data set was calibrated by all four estimation approaches (i.e., IBA/MULTI and SEM/XPD fully crossed).

x. The number of quadrature points used for approximation of the likelihood function was set to be15 points ranging between -4.0 and 4.0 (same for each latent factor). The number of quadrature points adopted was lower than the 19 points suggested in previous research (Cai, 2008; Paek & Cai, 2014). However, the integration range used in the current study was also adjusted accordingly to maintain comparable quadrature bin width.

xi. The user-specified maximum number of allowed EM cycles was fixed to 5000 (informed by pilot simulations) given the high dimensionality of the IBA models implemented in the current study.

xii. The maximum number of M-step iterations was 100 within each EM cycle, which was flexMIRT software's (Cai, 2015) default setting.

xiii. The E-step convergence criterion was adjusted to 1e-3 (versus the 1e-4 default), and the M-step convergence criterion was also adjusted to 1e-5 (versus the 1e-7 default). These convergence criteria were informed by email conversations with Dr. Li Cai who is the author of flexMIRT (Cai, 2015). As an expert on IBA modeling and DIF analysis, Dr. Cai has done extensive research in both of those areas (e.g., Cai, 2010; Cai et al., 2011; Woods et al., 2013).

xiv. The updated SEM procedure (Tian et al. 2013) was adopted (flexMIRT's default) for the two methods that utilized the SEM procedure, and the convergence criterion for the SEM algorithm was adjusted to 5e-3 to save computation time. This convergence criterion was reasonable (previous research suggested to use the square root of the original E-step criterion as the convergence criterion for the SEM procedure; Cai, 2008; Meng & Rubin, 1991), even though it was less stringent than the default 1e-3 convergence criterion found in flexMIRT.

Random variables. Several aspects of the current Monte Carlo study varied but were not manipulated. The decision to randomly simulate these variables was based on the fact that they are typically unknown factors in applied studies. Therefore, the process of drawing random values for these factors more or less mimicked the reality.

The proportion of DIF items per time point was randomly decided based on a uniform draw between 10% and 50% of items (i.e., between one and five items), and these proportions were informed by research found in the DIF literature (e.g., Wang, Shih, & Sun, 2012). The slopes and intercepts for each item were randomly generated, as well as DIF patterns (uniform versus nonuniform) and DIF magnitudes (see the Data Generation section for details).

Miscellaneous. Other aspects of the current study not presented in the previous sections:

i. The starting seed value of the random number generator in R (R core team, 2015) for the entire simulation was set to 20151017.

ii. The data-generating seed values in flexMIRT (Cai, 2015) were set to equal 300000 plus the current replication number (e.g., the number 300001 was used as the data-generating seed for the first replication).

Data Generation

The parameters of choice for this section were primarily informed by DIF studies mentioned previously (especially the ones which examined graded response items), although some minor adjustments were made (e.g., the DIF effect in the threshold is a uniform random draw between -0.7 and -0.3, instead of being manipulated as a factor with fixed levels at 0.3, 0.5 and 0.7; see Woods et al., 2013).

Primary factors. The mean and SD of the primary factor at Time-1 were fixed at 0 and 1 respectively. For Time-2, the primary factor mean and SD was simulated to be 0.5 and 1 respectively. For Time-3, the primary factor mean and SD was simulated to be 0.8 and 1 respectively. All primary factors were normally distributed, and they were assumed to be multivariate normal together with the specific factors described in the following section.

The correlation between primary factors followed a simplex structure where the correlation between two consecutive time points was 0.3, and the correlation between primary factors at Time-1 and Time-3 was calculated as $0.3 \times 0.3 = 0.09$. Implementations of the simplex structure for modeling factor correlations are common in the longitudinal measurement invariance literature (see Short, 2014).

Specific factors. All specific factors were fixed to be standard normal to avoid unnecessary variability of the resulting data. The total number of specific factors was equal to the test length (i.e., 10 specific factors for all replications throughout the study), and each specific factor was loaded by a triplet consisting of the same item taken across three time points. Note that each specific factor was permitted to covary neither with the primary factors nor with any other specific factors, in order for flexMIRT to implement dimension reduction during the estimation process.

Item slopes (primary and specific). Informed by past DIF research (e.g., Woods et al., 2013), item slopes on the primary factor at Time-1 were generated from a truncated normal distribution with $a_{j_P} \sim N(1.7, 0.6) \in [1.5, 4.0]$, and the primary slopes for subsequent time points remained the same as Time-1 for all DIF-free items.

Item slopes on the specific factors at Time-1 and Time-3 were randomly generated from a uniform distribution with $a_{j_S} \sim U(0.3, 0.7)$, and the specific slopes at Time-2 were randomly generated from a uniform distribution with $a_{j_S} \sim U(1.3, 1.7)$. Combining with values for the primary slopes, the slopes for the specific factors could also be translated to factor loadings in the factor analytic context using the equation

$$\lambda_{j_S} = \frac{a_{j_S}}{\sqrt{1+(a_{j_P}^2+a_{j_S}^2)}}, \quad (47)$$

which was derived from formulae 57 and 58 found in Hill (2006).

In contrast, the specific slope parameters in Hill (2006) were generated from a uniform distribution between 0.5 and 1.5, and they were fixed equal for both time points in that study (for model identification). For the current study, however, large specific slopes were generated only at Time-2 to simulate the situation where repeated items have a stronger relationship between each two consecutive time points than between Time-1 and Time-3.

Item intercepts. For each item, its first threshold at Time-1 for the primary factor was generated from a truncated normal distribution with $b_{j_P1} \sim N(-0.4,\ 0.9) \in [-2.5,\ 1.5]$. The distance between the first and the second thresholds is randomly drawn from a truncated normal distribution as $\Delta b_{j_P} \sim N(0.9,\ 0.4) \in [0.1,\ 0.625]$, and the same procedure repeats for generating the subsequent thresholds. The decision of using truncated normal distributions for generating the thresholds and their differences were primarily informed by Woods' (2009) review of applications (see Table 1 in Woods, 2009 for details). Necessary adjustments (e.g., limiting the maximum gap between two thresholds at 0.625) were made in order to prevent the reference group thresholds (i.e., Time-1 primary thresholds) from falling outside of the integration range between -4 and 4 (see Study Conditions section, Fixed variables subsection, number ix for details). For all DIF-free items, their primary thresholds at subsequent time points remained the same as Time-1.

All thresholds were then converted to intercepts using the formula

$$d_j = -a_{j_P} * b_{j_P}\ . \quad (48)$$

Note that, for each item, there could be only one set of intercepts, even though there were separate sets of thresholds with respect to primary and specific factors. By rearranging the above formula, we could obtain the thresholds for the specific factors using the intercepts and specific slopes, but the specific thresholds were of no interest for the current simulation.

DIF effects. For DIF items, their primary slopes at each subsequent time point varied by a uniform random draw between -0.7 and 0.7 (implemented sequentially for Time-2 and Time-3). The primary thresholds of a DIF item at Time-2 and Time-3 were always simulated to be smaller than the thresholds at Time-1. DIF magnitudes for the primary thresholds of each DIF item were randomly drawn from a uniform distribution between -0.7 and -0.3 (also implemented sequentially for Time-2 and Time-3). In other words, the probabilities of endorsing higher categories increased with lowered thresholds at Time-2 (and even lower thresholds at Time-3) compared to Time-1. The pattern of threshold changes remained consistent within each item (i.e., for a DIF item at a given time point, all thresholds changed in the same direction with the same magnitude). The corresponding intercepts of the DIF items were then calculated based on values of the shifted primary slopes and thresholds.

Procedure

The data generation process, the unidimensional MULTI model estimation, the modified Wald test for multiple-groups, the multidimensional IBA model calibration, and the SEM and XPD standard error approximation procedures were all implemented in flexMIRT version 3 (Cai, 2015) using three office computers located on the University of Kansas Lawrence campus (each equipped with Intel Core i7 860 2.8GHz quad-core processors and 4 GB of RAM). The compilation of flexMIRT scripts and the summarization of simulation results were automated using R version 3 (R Core Team, 2015).

The modified Wald test, however, was not automated in flexMIRT for analyzing results obtained from IBA-based estimation approach at the time of the current study (the modified Wald test option in flexMIRT was built-in only for comparisons of multiple independent groups). Therefore, the modified Wald test for longitudinal DIF detection was programmed in R by the author, utilizing item parameter estimates and the standard error covariance matrix (estimated via either SEM or XPD procedures) obtained from the IBA model calibrations. Power and Type I error associated with both the omnibus DIF test (which accounted for all contrasts in a contrast matrix **C)** and pairwise comparisons (which compared Time-2 and Time-3 respectively to Time-1) were evaluated.

Also note that the multiple-group Wald test in flexMIRT conducts separate, though simultaneous, group comparisons for each contrast in the **C** matrix. Hence, an additional R function was written to implement the omnibus multiple-group DIF test using item parameters and standard errors acquired from flexMIRT output, in order to maintain study-wise consistency.

Evaluated Outcomes

All the outcome measures presented below were calculated separately for each of the two anchoring conditions (i.e., the one-anchor and two-anchor conditions).

All measures of parameter estimation bias were calculated separately for either the IBA or MULTI estimation approach, unless otherwise stated. The choice of standard error estimation procedure (SEM or XPD) has no effect on the parameter estimation process because the error covariance matrix is calculated after the parameters are estimated. For instance, IBA-SEM and IBA-XPD will yield the same set of model parameters (including item parameters and latent distributions) despite having different standard errors. Thus, calculating bias for both IBA-SEM

and IBA-XPD would be repetitive. In addition, only the replications with convergent parameter estimates were considered in evaluations of estimation bias.

All four methods (IBA-SEM, IBA-XPD, MULTI-SEM, and MULTI-XPD) were compared in terms of statistical power and Type I error associated with the subsequent DIF analysis using the modified Wald test. In addition, for methods that utilize the SEM procedure, only the replications with both convergent parameter estimates and convergent standard error solutions were considered. A convergence criterion is not needed for XPD because it is not an iterative procedure.

Bias. Two measures of bias were computed for the estimated model parameters in the current study. The root mean squared error (RMSE) statistic was used to gauge the averaged distance between a set of estimates and their true values. The mean absolute percentage error (MAPE) statistic was used to quantify the averaged percentage distance (converted to absolute values) between a set of estimates and their true values. In each of the following subsections, equations for the calculation of RMSE and MAPE are provided.

Latent mean estimation. The RMSEs of latent mean estimates were calculated for Time-2 (true mean was 0.5) and Time-3 (true mean was 0.8) respectively as

$$RMSE_{\hat{\theta}_{T2}} = \sqrt{\frac{\sum_{rep=1}^{REPS}\left(\hat{\theta}_{T2_{rep}} - 0.5\right)^2}{REPS}}, \tag{49}$$

and

$$RMSE_{\hat{\theta}_{T3}} = \sqrt{\frac{\sum_{rep=1}^{REPS}\left(\hat{\theta}_{T3_{rep}} - 0.8\right)^2}{REPS}}, \tag{50}$$

where $\hat{\theta}_{T2_{rep}}$ and $\hat{\theta}_{T3_{rep}}$ are vectors that contain estimated factor means at Time-2 and Time-3 respectively across replications, and $REPS$ is the total number of replications (i.e., 500 for each of the two anchoring conditions if fully converged).

The MAPEs of latent mean estimates at Time-2 and Time-3 were also calculated using

$$MAPE_{\hat{\theta}_{T2}} = \frac{\sum_{rep=1}^{REPS} \left| \frac{\hat{\theta}_{T2_{rep}} - 0.5}{0.5} \right|}{REPS}, \tag{51}$$

and

$$MAPE_{\hat{\theta}_{T3}} = \frac{\sum_{rep=1}^{REPS} \left| \frac{\hat{\theta}_{T3_{rep}} - 0.8}{0.8} \right|}{REPS}, \tag{52}$$

where all components and symbols have the same interpretations as in the equations for calculating the RMSEs.

Latent SD estimation. The RMSEs of latent SD estimates at different time points were calculated similarly using

$$RMSE_{\widehat{SD}_{T2}} = \sqrt{\frac{\sum_{rep=1}^{REPS} \left(\sqrt{\widehat{SD^2}_{T2_{rep}}} - 1 \right)^2}{REPS}}, \tag{53}$$

and

$$RMSE_{\widehat{SD}_{T3}} = \sqrt{\frac{\sum_{rep=1}^{REPS} \left(\sqrt{\widehat{SD^2}_{T3_{rep}}} - 1 \right)^2}{REPS}}, \tag{54}$$

where $\widehat{SD^2}_{T1_{rep}}$ and $\widehat{SD^2}_{T2_{rep}}$ are estimated factor variances at Time-2 and Time-3, and 1 is the true value of simulated SD for both time points.

The MAPEs were computed as

$$MAPE_{\widehat{SD}_{T2}} = \frac{\sum_{rep=1}^{REPS} \left| \sqrt{\widehat{SD^2}_{T2_{rep}}} - 1 \right|}{REPS}, \tag{55}$$

and

$$MAPE_{\widehat{SD}_{T3}} = \frac{\sum_{rep=1}^{REPS}\left|\sqrt{\widehat{SD^2}_{T3rep}}-1\right|}{REPS}. \quad (56)$$

The denominators inside the parentheses dropped out in the equations given that they both equal 1.

Latent correlation estimation. The RMSEs of latent correlation estimates were calculated for all three possible correlations between each pair of primary factors (only applicable to IBA models). The RMSE of the estimated factor correlation between Time-1 and Time-2, for example, was computed as

$$RMSE_{\hat{r}_{T1T2}} = \sqrt{\frac{\sum_{rep=1}^{REPS}\left(\frac{\widehat{COV}_{T_1T_2rep}}{\sqrt{\widehat{SD^2}_{T1rep}\widehat{SD^2}_{T2rep}}}-0.3\right)^2}{REPS}}, \quad (57)$$

where $\widehat{COV}_{T1T2_{rep}}$ is the covariance between primary factors at Time-1 and Time-2, and $\widehat{SD^2}_{T1_{rep}}$ and $\widehat{SD^2}_{T2_{rep}}$ are aforementioned variance estimates. And the corresponding MAPE was computed as

$$MAPE_{\hat{r}_{T1T2}} = \frac{\sum_{rep=1}^{REPS}\left|\left(\frac{\widehat{COV}_{T_1T_2rep}}{\sqrt{\widehat{SD^2}_{T1rep}\widehat{SD^2}_{T2rep}}}-0.3\right)\left(\frac{10}{3}\right)\right|}{REPS}. \quad (58)$$

The conversion from covariance to correlation was necessary due to the fact that the estimated primary factor variances at Time-2 and Time-3 were not always equal to 1, thus the covariance from the model output could not be directly interpreted as correlations between factors.

Similar formulae were used for calculating the RMSE and MAPE of the estimated factor correlation between Time-2 and Time-3 with changes only in the subscripts, given that the simulated correlation at each two consecutive time points was always 0.3.

The equations for calculating the RMSE and MAPE of the estimated factor correlation between Time-1 and Time-3 replaces 0.3 in equation (57) by 0.09 (and therefore replace $\frac{10}{3}$ in equation 58 by $\frac{100}{9}$), which was the true generating value for the correlation between Time-1 and Time-3.

Primary slope and intercept estimation. The RMSEs and MAPEs of primary slope parameter estimates for non-anchors (including DIF-free tested items and true DIF items) were calculated separately for each time point within each replication. The RMSEs and MAPEs of primary slope parameter estimates for anchors were calculated at the first time point only, because anchor parameters were equated across time. The general equations for the RMSE and MAPE calculations were

$$RMSE_{\hat{a}_P} = \sqrt{\frac{\sum_{j=1}^{J}\left(\hat{a}_{j_P} - a_{j_P}\right)^2}{J}}, \tag{59}$$

and

$$MAPE_{\hat{a}_P} = \frac{\sum_{j=1}^{J}\left|\frac{\hat{a}_{j_P} - a_{j_P}}{a_{j_P}}\right|}{J}. \tag{60}$$

In both equations, $\hat{a}_{j_P}$ is the estimated primary slope for item *j* at a given time point, and a_{j_P} is the data-generating true primary slope, and *J* is the number of compared items. Regarding the index *J*, it could take a value of 8 or 9 for non-anchors within each time point, and a value of 1 or 2 for anchors.

Similarly for each non-anchor intercept estimate within each replication, individual RMSEs and MAPEs were also calculated separately at each time point. For anchors, only bias within Time-1 was evaluated because their intercepts were also equated across time. The equations used for the calculations of the RMSE and MAPE were

$$RMSE_{\hat{d}_v} = \sqrt{\frac{\sum_{j=1}^{J}(\hat{d}_{v,j} - d_{v,j})^2}{J}}, \tag{61}$$

and

$$MAPE_{\hat{d}_v} = \frac{\sum_{j=1}^{J}\left|\frac{\hat{d}_{v,j} - d_{v,j}}{d_{v,j}}\right|}{J}. \tag{62}$$

In both equations, $\hat{d}_{v,j}$ is the estimate for the v^{th} intercept ($v \in \{1, 2, 3, 4\}$) for item j at a given time point, $d_{v,j}$ is the data-generating true intercept, and their differences were averaged over J number of compared items. For non-anchors, the value of J varied depending on how many items had categories collapsed at a given time point. For anchors, the value of J could only take on a value of 1 or 2.

Specific slope estimation. Bias in the estimation of specific slopes under the IBA models were also calculated separately for each specific factor within each replication using

$$RMSE_{\hat{a}_S} = \sqrt{\frac{\sum_{t=1}^{3}(\hat{a}_{S_t} - a_{S_t})^2}{3}}, \tag{63}$$

and

$$MAPE_{\hat{a}_S} = \frac{\sum_{t=1}^{3}\left|\frac{\hat{a}_{S_t} - a_{S_t}}{a_{S_t}}\right|}{3}, \tag{64}$$

where $\hat{a}_{S_t}$ is the estimated specific slope at time t (where $t \in \{1, 2, 3\}$) within a replication, and a_{S_t} is the true specific slope. Both the sum of squared differences (for RMSE) and absolute percentage differences (for MAPE) were averaged over the three time points.

Note that items within a given time point all loaded on different specific factors, but each cross-time triplet loaded on the same specific factor. Therefore, the bias measures of specific slopes were calculated for each triplet of items within a replication, instead of being calculated separately for each time point. Consequently, there was no need to differentiate anchors from

non-anchors, because every triplet of anchors always loaded on the same specific factor throughout the study.

Statistical power and Type I error. True positive rate (statistical power) for each replication was calculated as the proportion of true DIF items being flagged as having DIF. The false positive rate (Type I error) for each replication was calculated as the proportion of true DIF-free items being mistakenly flagged as having DIF.

Throughout the current study, DIF in both the slope and all intercepts were jointly tested for DIF using equation (9). Tests of individual parameters (e.g., slope only) were not implemented. Therefore, only the joint test statistic was evaluated for power and Type I error.

Computation time. Computation time (in seconds) associated with each step of the parameter and error covariance matrix estimation was collected from flexMIRT output. Comparisons of computation time of all four methods were conducted to demonstrate the differences between IBA and MULTI in terms of model estimation time and between SEM and XPD in terms of standard error calculation time.

Results

Category Collapse

No category collapsing occurred among the simulated anchor items throughout the current study. Only 4.8% of the 1000 total replications had one studied item (non-anchor) with category collapses, and the minimum number of categories observed for any items across replications was two (so these items were treated as 2PL items by flexMIRT). A total of 46 replications (out of the 48 cases) had category collapsing on DIF items. DIF items accounted for the majority of category collapsing due to their elevated thresholds working in conjunction with the higher latent factor mean, especially at Time-3.

In order to implement the modified Wald test, the number of categories for an item was forced to be equal across all three time points (e.g., forced category collapsing of an item at Time-1 and Time-2, if it had only two categories at Time-3). Prior to model calibration, category matching was accomplished by recoding the responses for items that showed differing categories between time points. For example, responses '3' through '5' of an item were all coded as response '3' when the item was matched to its three-category counterpart at a different time point. Had there been anchor items with collapsed categories, this recoding procedure would have been unnecessary because the anchors were not required to have the same number of parameters across time for DIF analysis purposes.

In addition, all 48 replications with category collapsing converged during the estimation process under both IBA and MULTI. Out of the 11 replications with non-convergent SEM error estimation under the IBA models, two cases had items with collapsed categories. Out of the 885 non-convergent SEM under the MULTI models, 41 cases had items with collapsed categories. See the following section for more details regarding convergence.

Convergence of Model Calibration and Error Estimation

The overall model calibration convergence rate was computed, for either the IBA or the MULTI estimation approaches, as the proportion of replications converged under the user-specified convergence criteria (E-step at 0.001 and M-step at 0.00001) and within the maximum number of allowed EM cycles (i.e., 5000). Specifically for the two approaches that implemented the SEM procedure for error covariance estimation, the convergence rates of the SEM algorithm were also monitored and recorded.

Overall, 10 replications did not converge during estimation of the multidimensional IBA model, whereas estimation of the unidimensional MULTI model fully converged for the 1000 replications.

During the standard error calculation step using the SEM procedure, when the data were fit using IBA, 11 out of the 1000 replications did not converge (within the maximum allowed iterative SEM cycles set by flexMIRT's default). These cases with non-convergent standard errors did not overlap with any of the cases with non-convergent parameter estimates. The causes of this phenomenon (i.e., the mutually exclusive non-convergence issues) are unknown and require further investigation in future studies.

When the SEM procedure was implemented following the MULTI estimation approach, 885 out of the 1000 replications did not converge on a solution for the error covariance matrix. The astonishing non-convergence problem associated with the MULTI-SEM method was no surprise since the calibrated model was misspecified and independence assumptions were violated. An interesting fact was that the MULTI model had no convergence problems while estimating the parameters, despite the severely low non-convergence rate when the SEM procedure was applied for standard error computation. In an evaluation of several standard error calculation procedures (including SEM) conducted by Paek and Cai (2014), SEM had no convergence problems when the model for parameter estimation was correctly specified. Thus, the misapplication of the MULTI-SEM method to longitudinal data could even be informative in a sense that, a model is likely misspecified (e.g., violation of independence assumptions) when convergent parameter estimates are obtained along with non-convergent standard errors.

Even though the XPD procedure does not involve an iterative process (hence no convergence criterion required during standard error calculation), improper standard errors could

be produced by the XPD procedure due to unknown reasons. For example, post simulation analysis showed that one replication calibrated by the IBA-XPD method produced a standard error as large as 147312 times its corresponding parameter estimate. Throughout the study, there were two instances of such unusual standard error values when XPD was used for standard error calculation following the IBA-based parameter estimation, which prevented the modified Wald test from being implemented due to singularity of the dispersion matrix (so that the sigma matrix could not be inverted). Therefore, these two replications were removed from all subsequent analyses.

Bias

All RMSE and MAPE calculations excluded non-convergent parameter estimates.

The RMSEs and MAPEs associated with either IBA or MULTI were calculated separately for each anchoring condition as measures of estimation bias at each time point. For factor level biases, there was a single set of RMSE and MAPE values associated with each method under each anchoring condition.

For item level biases, condition-wise RMSE and MAPE values were calculated at each time point for each primary parameter (one slope and four intercepts) within each replication. Both bias measures associated with specific slope estimation were also calculated for each replication when IBA was implemented.

Latent mean estimation. The raw values of estimated latent means were plotted in *Figure 2* and *Figure 3*. As expected, IBA showed a dense distribution centered on the true value of 0.5, whereas MULTI showed more variability, under both one-anchor and two-anchor conditions at Time-2. At Time-3, MULTI was on a par with IBA in terms of mean estimation accuracy, especially under the one-anchor condition where the two distributions almost entirely

overlapped and centered on the true value of 0.8. To examine the estimation bias in more detail, RMSE and MAPE measures were calculated and plotted.

As shown in *Figure 4* and *Figure 5* in Appendix A, the factor means estimated by IBA were overall less biased than those estimated by MULTI at corresponding time points under both anchoring conditions. The RMSEs and MAPEs associated with MULTI's estimation of the latent mean at Time-2 were noticeably higher (about twice the bias associated with the IBA counterparts). The large bias associated with MULTI at Time-2 was expected since its unidimensional model did not account for the factor covariance/correlation between Time-1 and Time-2 (simulated as 0.3). Even though a small correlation (0.09) was also simulated between Time-1 and Time-3, MULTI was not affected by the weak correlation between the two time points.

The RMSEs agreed with the MAPEs, even though all Time-2 MAPEs seemed inflated. This was due to the fact that MAPEs were scaled differently according to the true latent means (0.5 at Time-2 and 0.8 at Time-3). An average 0.1-unit non-directional discrepancy between the estimated mean and the true mean, for example, would be translated to 20% bias in the percentage scale for Time-2 but 12.5% bias for Time-3.

Latent SD estimation. The raw values of estimated latent SDs were plotted in *Figure 6* and *Figure 7*. When a moderate correlation is present between Time-1 (the reference time point) and Time-2, MULTI tended to underestimate the SD at Time-2 under both anchoring conditions. In contrast, IBA showed considerably less bias in terms of SD estimation at Time-2, and the densities under both anchoring conditions centered on 0.98 (the true value was 1.0). At Time-3, the performance of these two methods was comparable, although IBA still outperformed MULTI.

To examine the estimation bias in more detail, RMSE and MAPE measures were calculated and plotted.

As shown in *Figure 8*, MULTI always had a higher RMSE than IBA under corresponding conditions, except at Time-3 with only one anchor. MULTI had trouble estimating the SDs for Time-2 under both anchoring conditions (each was about twice the corresponding RMSEs associated with IBA). In *Figure 9*, the MAPEs showed a similar pattern regarding estimation bias in latent SDs between the two methods. The percentage bias of SD estimation was on the same scale (the true SD was always 1.0) so the MAPE measures could be directly compared.

Latent correlation estimation. For the IBA method, it was also necessary to examine the bias associated with latent correlation estimation. Density plots for the raw values of the correlation estimates were plotted in *Figure 10*. IBA showed little bias in terms of the estimated correlations between time points with all densities centered on the true values (i.e., 0.3 between two consecutive time points, and 0.09 between Time-1 and Time-3).

In accordance with the density plots, the RMSEs plotted in *Figure 11* showed little bias in the correlation estimation, even though the correlation estimates between Time-1 and Time-3 showed a higher percentage bias in the MAPE plot. Such high percentage biases should not be too concerning in practice, given that the biases found in the RMSE plot were very consistent across the board.

Primary slope and intercept estimation. For anchors, RMSEs and MAPEs were computed at Time-1 only for each replication, because anchor parameters were equated across time. For non-anchors, bias measures were computed separately at each time point within each replication.

Anchor parameter estimation. In order to maintain consistency regarding the scale of the ordinate, small portions of observations that had RMSEs greater than 1.0 (or 100% in terms of MAPE) were not shown on the boxplots found in *Figure 12* and *Figure 13*.

Under the one-anchor condition, in terms of RMSE (first column of *Figure 12*), none of the upper whiskers of the IBA-estimated parameters exceeded 0.4, whereas the RMSEs associated with MULTI had upper whiskers that extended as high as close to 0.8. In terms of MAPE (second column of *Figure 12*), IBA also showed lower percentage biases with all medians below 10%.

Under the two-anchor condition, in terms of RMSE (first column of *Figure 13*), all boxplots associated with IBA were more condensed than their one-anchor counterparts, whereas all boxplots associated with MULTI showed elevated median levels and similar variability compared to their one-anchor counterparts. A similar pattern was found for MAPE (second column of *Figure 13*) with MULTI showing worse percentage bias compared to the one-anchor condition.

Overall, IBA showed less bias than MULTI in terms of anchor parameter estimation. IBA-estimated anchor parameters displayed less bias under the two-anchor condition than under the one-anchor condition, whereas the addition of an extra anchor was detrimental to MULTI. The observed pattern was expected given that MULTI ignored the factor and item covariations across time. Moreover, these patterns were consistent with findings from the latent mean and SD estimation bias (IBA had more accurate and reliable latent mean and SD estimates), because anchors were used to estimate the latent distributions at Time-2 and Time-3.

Non-anchor parameter estimation at Time-1. In order to maintain consistency regarding the scale of the ordinate, small portions of observations that had RMSEs greater than 1.0 (or 100% in terms of MAPE) were not shown on the boxplots found in *Figure 14* and *Figure 15*.

Under the one-anchor condition, as shown in *Figure 14*, the medians of the set of boxplots associated respectively with RMSEs and MAPEs were about the same between IBA and MULTI. Nonetheless, IBA showed more variability in estimation bias for non-anchor items. Both IBA and MULTI showed minimal percentage bias (i.e., MAPE) associated with the slope parameter estimation.

Under the two-anchor condition, as shown in *Figure 15*, nothing had noticeably changed for both IBA and MULTI. The overall pattern of RMSE and MAPE, regardless of the type of method, were quite similar between the one-anchor and two-anchor conditions. Both IBA and MULTI benefited slightly from the inclusion of an additional anchor.

The observed pattern at Time-1was not surprising, even though MULTI seemed to outperform IBA in terms of non-anchor item parameter estimation. On one hand, MULTI had accurate estimates because Time-1 was the reference point, and therefore the estimation of Time-1 non-anchor parameters had nothing to do with items at other time points (they were not equated across time points the same way as anchors). On the other hand, as a multidimensional model, IBA had many more parameters to estimate, and therefore it was reasonable to expect IBA to encounter more "local maxima" problems in the multidimensional space when implementing EM-MML. Future research could consider a hybrid approach, where MULTI-estimated non-anchor item parameters from the reference time point can be used for the actual IBA model.

Non-anchor parameter estimation at Time-2. Consistency regarding the scale of the ordinate could not be maintained for Time-2 boxplots, due to the unexpected large estimation bias associated with IBA. All boxplots, except the upper left ones in both *Figure 16* and *Figure 17*, used an ordinate scale between 0 and 5, with small proportions of outliers not shown on the plots. The upper left plots (i.e., the RMSEs associated with IBA) in both *Figure 16* and *Figure 17* used an ordinate scale between 0 and 15 so that the upper whiskers could be visible on the plots.

All else being equal, there was no noticeable difference between the one-anchor condition and two-anchor condition in terms of estimation bias for non-anchor items at Time-2. However, IBA had surprisingly biased non-anchor item parameters at Time-2, compared to MULTI. In terms of RMSE, the boxplots associated with IBA had much wider ranges and higher medians. Even though smaller differences between IBA and MULTI were found in terms of MAPE especially when looking at the medians, the IBA-estimated non-anchor item parameters were still biased as large as 300%. Nevertheless, performance of MULTI at Time-2 was also found to be much worse than its performance at Time-1, with most of the MAPE medians being above 45% and upper whiskers reaching close to (or higher than) 100%.

Since IBA had more accurate anchor parameter estimates and better latent distribution estimates, one would expect IBA to behave similarly when estimating non-anchor parameters at Time-2. However, this was not the case according to *Figure 16* and *Figure 17*. One possible explanation is that IBA had one or two extremely biased Time-2 parameter estimates within each replication, which distorted the RMSEs and MAPEs associated with IBA at Time-2 when the average of within time point bias was taken. In order to find out the reason for the unexpectedly high RMSEs and MAPEs associated with IBA at Time-2, additional steps were taken to

investigate the non-averaged absolute percentage bias across all non-anchor parameters estimated by IBA in the current study.

As shown in the first row of Figure 18, the average percentage biases of all IBA-estimated non-anchor item parameters had a few extremely large values (e.g., more than 400000% bias under the one-anchor condition). However, such extreme cases were sporadic beyond the cutoff of 5 on the ordinate scale (or 500% bias). When the boxplots were magnified to take a closer look at the absolute percentage bias on a scale between 0 and 5, the IBA-estimated parameters had all median biases (in terms of absolute percentage bias) below 0.3 (or 30% bias, as indicated by the red dotted lines). Therefore, it was evident that the previously observed large RMSEs and MAPEs associated with IBA were distorted due to the existence of those extremely biased outliers.

Moreover, the subsequent DIF analysis might not be substantially affected by these outliers, because each item is tested for DIF one at a time rather than after an averaging process within a time point such as computing the RMSE or MAPE for the estimates. In addition, the DIF analysis also takes the standard errors into account when calculating the test statistic. Hence, a biased parameter estimate does not necessarily result in reduced power or inflated Type I error during DIF analysis. See the '**Statistical Power and Type I Error**' section for more details regarding the power and Type I error associated with IBA-based methods.

Non-anchor parameter estimation at Time-3. In order to maintain consistency regarding the scale of the ordinate, small portions of observations that had RMSEs greater than 1.0 (or 100% in terms of MAPE) were not shown on the boxplots found in *Figure 19* and *Figure 20*.

As expected, estimation bias at Time-3 was comparable to what was found at Time-1, despite small inflations at Time-3. On one hand, the simulated factor correlation between the

two time points (i.e., 0.09) was very small. On the other hand, the simulated item covariations (captured by the specific factors) were very small, with simulated specific slopes for the two time points ranging between 0.3 and 0.7.

Overall, MULTI still had more condensed bias measures than IBA, and both methods benefited from the inclusion of an additional anchor. For both IBA and MULTI, the median MAPEs were all under 20%, and the median bias of slope estimates fell below 10%.

Specific slope estimation. Bias in the estimation of specific slopes (estimated under IBA models only) was also measured using RMSE and MAPE. In order to maintain consistency regarding the scale of the ordinate, small portions of observations that had RMSEs greater than 3.0 (or 300% in terms of MAPE) were not shown on the boxplots found in *Figure 21* and *Figure 22*.

As shown in *Figure 21* and *Figure 22*, the median RMSEs were all below 0.5 and the median MAPEs were all below 50%, under both anchoring conditions. Under the one-anchor condition, estimation bias associated with specific factor S1 was much lower than the other specific factors. The inflated bias associated with the other specific factors might partly be due to the inaccurate primary parameters estimated by IBA at Time-2. Bias in estimation of S1 slopes were not affected because IBA-estimated anchor parameters (equated across time) were found to be accurate. There was no noticeable difference in terms of estimation bias between the two-anchor condition and the one-anchor condition, except that the bias associated with S2 (which connected the second anchor across time) was lower under the two-anchor condition.

Statistical Power and Type I Error

True positive rate (statistical power) for each replication was calculated as the proportion of true DIF items being flagged as having DIF. False positive rate (Type I error) for each

replication was calculated as the proportion of true DIF-free items being mistakenly flagged as having DIF.

Power and Type I error were only calculated for convergent replications. For the two methods using SEM as the standard error approximation procedure, only replications with convergent standard errors were taken into consideration when evaluating power and Type I error. In addition, power and Type I error were calculated separately for the omnibus DIF test (all three time points compared simultaneously using a contrast matrix) and also for each of the two pairwise DIF tests (which compared Time-2 and Time-3 respectively to Time-1).

Generally speaking, all four methods showed very high power in DIF detection. In terms of Type I error, IBA-based methods outperformed MULTI-based methods, although IBA-XPD was the only method that had Type I error controlled below .05.

Omnibus DIF test. In general, all four methods (IBA-SEM, IBA-XPD, MULTI-SEM, and MULTI-XPD) had unanimously high power when detecting omnibus DIF. As shown in the first row of *Figure 23*, almost all replications had a 100% true positive rate in terms of detecting DIF items. This pattern of high statistical power was not surprising given the large sample size (i.e., 1000).

However, as shown in the second row of *Figure* 23, Type I errors associated with the omnibus DIF test followed a drastically different pattern depending on the method. The IBA-based methods had markedly lower Type I errors than MULTI-based methods. Such finding was expected because MULTI-based methods used misspecified unidimensional models to estimate the data-generating multidimensional factor structure, and therefore it could not distinguish latent-level changes from item-level changes. Furthermore, IBA-XPD was the only method had Type I error maintained below .05 (indicated by the red dotted line) across the majority of

converged replications, especially under the two-anchor condition where IBA-XPD had almost no Type I error across converged replications.

Pairwise DIF tests. Power and Type I errors were also evaluated for each of the two pairwise DIF tests across converged replications. Boxplots associated with each pairwise DIF test appear in *Figure 24* (Time-1 and Time-2 compared) and *Figure 25* (Time-1 and Time-3 compared). No comparison between Time-2 and Time-3 was conducted because the omnibus test had a contrast matrix comparing both Time-2 and Time-3 to Time-1 (the reference time point).

Time-1 and Time-2 compared. As shown in the first row of *Figure 24*, power associated with the pairwise comparison between Time-1 and Time-2 was maintained at 100% for IBA-SEM, MULTI-SEM, and MULTI-XPD methods, under both anchoring conditions. Although IBA-XPD's power was noticeably lower than the other three methods under both anchoring conditions, its power was maintained at or above 80% across the majority of replications. Type I errors associated with the same pairwise comparison showed a pattern similar to the omnibus test results. IBA-XPD was still the only method that had Type I error controlled within .05 across the majority of replications, especially under the two-anchor condition.

As mentioned previously, IBA produced a few extremely biased outliers when estimating parameters for non-anchor items at Time-2. However, items were tested for DIF individually, and therefore DIF test results should not be substantially affected by outliers with extreme biases. Despite the fact that MULTI yielded more accurate non-anchor parameter estimates than IBA at Time-2 (see Bias section for details), IBA-SEM/XPD outperformed MULTI-based methods in the current pairwise DIF test (see *Figure 24*).

Time-1 and Time-3 compared. As shown in the first row of *Figure 25*, there was 100% statistical power for the pairwise comparison between Time-1 and Time-3 for all methods. Meanwhile, Type I error was unanimously low, with the majority of replications across all methods showing Type I errors below .05. It was not surprising that the MULTI-based methods had well-controlled Type I error associated with the pairwise comparison between Time-1 and Time-3. The simulated factor correlation between Time-1 and Time-3 (i.e., 0.09) was very small, as were the simulated specific slopes for the two time points (ranged between 0.3 and 0.5). Therefore, results from MULTI-based methods comparing Time-1 and Time-3 in the current study were expected to be minimally affected by model misspecification.

Computation Time

The required computation time for EM algorithm convergence and standard error computation were calculated separately for the four methods under both anchoring conditions. The unit of time shown in *Figure 26* was "hours" for IBA-based methods and "minutes" for MULTI-based methods, in order to put respective methods on more meaningful scales of time.

As shown in the upper left boxplot of *Figure 26*, the majority of replications using the IBA approach required at least 15 hours until the EM algorithm converged on a solution. The time difference between IBA-SEM and IBA-XPD was negligible, given that the EM algorithm was only used for estimating the model parameters, and standard errors were calculated in a different procedure after EM convergence.

The upper right boxplot of *Figure 26* showed the number of hours required for each IBA-based method to calculate the standard errors for the parameters estimated via EM-MML. The XPD procedure almost always calculated the standard errors within 15 minutes, whereas the SEM procedure required at least four hours across the majority of the replications.

The EM convergence time required by MULTI-based methods was always less than six seconds, and the time required for standard error calculation was almost always within six seconds. However, the underlying model implemented by MULTI was unidimensional, and therefore it was inappropriate to use MULTI for longitudinal data, no matter how much time it saved compared to the IBA approach.

Empirical Example

Given satisfactory results from the simulation study, the proposed IBA-XPD method paired with the modified Wald test was also applied to the analysis of an empirical longitudinal data set for illustration purposes. The flexMIRT script used for modeling the data using the IBA approach appears in *Appendix C*, and the code for the automated R function used for calculating the Wald statistics is given in *Appendix D*.

Data Properties

The two-wave (panel two) data were collected by Zelinski and Kennison (2011) in 1994 and 1997, and the analyzed data set was downloaded from the ICPSR website under the name "Long Beach Longitudinal Study" and the case number ICPSR 26561.

One section of the surveys implemented in 1994 and 1997 collected elderly respondents' responses on the 64-item Memory Functioning Questionnaire (MFQ; Gilewski, Zelinski, & Schaie, 1990). The original MFQ questions were all Likert-type items with options ranging from 1 to 7, with the lowest option 1 indicating "having severe memory functioning problems", and the highest option 7 indicating "having no memory functioning problems". The actual wording of the options changed in a few subsections in the survey, but they generally had the same interpretation in the context of memory functioning.

For demonstration purposes, only 10 items related to "frequency of forgetting" within the original MFQ survey were analyzed in the current study. These items were found to be highly reliable, and they correlated highly with the original "frequency of forgetting" items within the MFQ survey (Zelinski & Gilewski, 2004). Furthermore, the first item on the 10-item subscale was designated as the time-invariant anchor. At the time of this writing, none of the published studies had properly evaluated the measurement invariance properties of these 10 items used across time. Hence, the "general rating" question (i.e., the first item out of the 10 questions) was chosen as the anchor based on face validity.

After matching respondents from both time points by their case IDs, there were 358 subjects who participated in the longitudinal study in both 1994 and 1997. However, some subjects skipped the MFQ questions entirely in either 1994 or 1997. Therefore, additional steps were taken to list-wise delete those cases. The resulting 328 subjects responded to all of the 10 questions at both time points.

Analysis Procedures

The two-wave longitudinal data of the 328 respondents were fitted to an IBA model with two primary factors and 10 specific factors. The standard errors were calculated using both the SEM and XPD procedures. The flexMIRT script used for modeling the data using the IBA model is given in *Appendix C*. The item parameters estimated by the IBA model are presented in *Table 4*.

In the latest version of flexMIRT (version 3.0RC; Cai, 2015), the option for conducting DIF tests cannot be used in conjunction with the calibration of an IBA model. Therefore, a user-programmed R function was implemented to automate the process of Wald statistic calculation for the comparison of IBA-estimated parameters. The source code for the user-programmed R

function is provided in *Appendix D*. It was designed to work with flexMIRT output files if the parameters and the error covariance matrix (obtained using either SEM or XPD) were saved as separate files (using the "SavePRM = YES;" and "SaveCOV = YES;" options when composing the flexMIRT script; see details regarding the flexMIRT script used for the current empirical example in *Appendix C*). In addition, the R function shown in *Appendix D* conducts both the omnibus and by-contrast (not limited to pairwise) DIF tests using the modified Wald test.

DIF Test Results

Because the empirical data consisted of only two time points, only the pairwise DIF test was conducted to compare the parameters estimated for the year of 1994 and the year of 1997. This could be considered the omnibus test, given that all the time points were compared in one step.

As shown in *Table 5*, none of the studied items showed significant DIF effects when parameters and standard errors were estimated using the IBA-XPD approach. Recall that in the simulation study, IBA-XPD paired with Wald had relatively lower statistical power when pairwise comparisons were conducted between two time points that had strong covariations. Therefore, the same data were reanalyzed using the IBA-SEM method, given its extremely low Type II error found previously. Nonetheless, IBA-SEM found no item with a significant DIF effect (see *Table 6* for details).

Discussion

The current simulation study proposes a new approach for detecting DIF effect in longitudinal contexts. Specifically, the proposed approach involves three steps: 1) Estimate model parameters via EM-MML using IBA models with anchor item parameters constrained equal across time; 2) Approximate standard errors using either SEM or XPD; and 3) Detect

cross-time DIF effects using the modified Wald test designed to compare multiple sets of parameters simultaneously (with appropriate parameter estimates and standard errors obtained from the preceding steps).

During the study, the multidimensional IBA modeling approach was compared to the traditional unidimensional MULTI approach in terms of convergence rate and model parameter estimation bias. Both the SEM and the XPD standard error computation procedures were also compared to each other, in terms of statistical power and Type I error associated with the subsequent DIF analysis carried out using the modified Wald test. Computation time required by all four methods (i.e., IBA-SEM, IBA-XPD, MULTI-SEM, and MULTI-XPD) was also evaluated.

Findings of the Current Simulation

As expected, IBA yielded much more accurate and reliable latent mean and SD estimates (for Time-2 and Time-3) than MULTI, when Time-1 was used as the reference time point (see *Figure 2* through *Figure 9* for details). This finding was consistent with Hill (2006) in which the bifactor approach was implemented for modeling simulated data of two time points, and small biases associated with latent distribution estimation was found under the bifactor approach. IBA's outstanding performance in latent distribution estimation was primarily attributed to its highly accurate anchor item parameter estimates (see *Figure 12* and *Figure 13* for details).

IBA suffered from severely biased non-anchor item parameter estimates at Time-2 (see *Figure 16* and *Figure 17*), even though further investigation found that the majority of estimates had reasonable absolute percentage bias, and extreme bias occurred only in a small proportion of estimates (see *Figure 18*). In Hill (2006), the bifactor approach for modeling longitudinal data was also found to produce less desirable parameter estimates compared to other competing

methods (e.g., the latent class analysis model). This finding, along with Hill's (2006), should be concerning if researchers plan to extend the bifactor approach for modeling even higher dimensional factor structures. With increased dimensionality, and the presence of substantial factor correlation and/or item covariation, IBA modeling via EM-MML might behave in an unpredictable way (and therefore has difficulty providing accurate parameter estimates).

Regarding power and Type I error, IBA-based methods outperformed MULTI-based methods as expected. IBA-XPD was the only method that maintained Type I error below .05 across the majority of replications for both the omnibus and pairwise DIF tests. Especially for the omnibus DIF test under the two-anchor condition, IBA-XPD almost always had Type I error close to nil while maintaining high statistical power (see *Figure 23*). Thus, the IBA-XPD method, paired with the modified Wald test, is suitable for detecting longitudinal DIF. All the other three methods struggled with having too much statistical power and inflated Type I error in DIF detection, even though IBA-SEM was noticeably better (Type I error controlled within .4 across the majority of replications) than the two MULTI-based approaches.

As to computation time, IBA models in general required more than 10 hours until EM-MML convergence, whereas MULTI models always required less than six seconds. The XPD procedure in general required much less computation time than the SEM procedure especially under the IBA models (see *Figure 26*).

Suggestions for Applied Research

Based on the foregoing discussions, the following three steps are recommended for cross-time DIF detection:

i. Fit appropriate IBA models to the longitudinal data set being studied, with known anchors equated across time points.

ii. Compute the standard errors using the XPD procedure.

iii. Calculate the Wald test statistics using the parameters and standard errors obtained from previous steps.

Both the model estimation and standard error computation steps can be carried out using flexMIRT (Cai, 2015), or any other IRT software that implements the IBA approach and the XPD procedure. For calculations of the Wald test statistics, researchers can either follow the steps detailed in Kim et al. (1995), or use the R function provided in Appendix D which works with output files saved from flexMIRT's calibration of the IBA models (the source code for the R function can be requested from the author via email at mian@ku.edu).

The recommended method, when it is carried out under conditions that are similar to those found in the current study, is expected to provide accurate latent-level parameter estimates, as well as high power and well-controlled Type I error associated with DIF tests. Even though non-anchor item parameter estimates were found to be less accurate in the simulation, additional steps can be taken to improve the item-level estimates especially for items showing no DIF effect. Researchers can first separate out DIF items using the recommended DIF detection method, and then include only DIF-free items (i.e., items with non-significant Wald test statistics) in a second run of the IBA model calibration to re-estimate the item parameters (equated across time points).

Nonetheless, readers should be aware that the recommended longitudinal DIF detection method might not work well under conditions that were not evaluated in the current study. Specifically, Paek and Cai (2014) found that the XPD procedure yielded noticeably more biased standard errors than SEM did under the bifactor structure, when the sample-size-to-test-length ratio was around 10. In general, the bias in standard error computation diminishes as the sample-

size-to-test-length ratio increases. Therefore, for researchers who intend to apply the IBA-XPD-Wald method to the analysis of empirical longitudinal data, it is recommended that the sample-size-to-test-length ratio is maintained above 40.

Limitations and Future Directions

Due to the extensive amount of time required by the IBA model estimation, and in order to complete the simulation in a timely manner, the current study only varied the proportion of anchors as the manipulated variable. Future research is needed to evaluate the performance of the longitudinal DIF method recommended by the current study (i.e., IBA-XPD paired with Wald) under various controlled simulation conditions (e.g., varying sample size and test length).

Also, throughout the current study, the E-step and M-step convergence criteria were set respectively at 0.001 and 0.00001, and the number of quadrature points used for numerical approximation during EM was adjusted to be 15 points ranging between -4.0 and 4.0. Previous research has mostly adopted the convention of using 19 points between -5.0 and 5.0 (Cai, 2008; Paek & Cai, 2014). In addition, the convergence criterion for the SEM procedure was also changed to 0.005, which is different from flexMIRT's default criterion of 0.001. Future investigations should shed light on whether increasing the number of quadrature points, and/or applying more stringent convergence criteria, would improve the performance of the modified Wald test using the parameters and standard errors estimated using IBA-SEM.

Furthermore, in the current study, anchor items remained as time-invariant across all time points, whereas DIF items were simulated to exhibit DIF effects at both Time-2 and Time-3. In reality, however, items that are invariant between two time points might show DIF effects at another time point and vice versa. Hence, future research could explore different patterns of DIF effects occurring across multiple time points.

Meanwhile, the item type used in the current study was also fixed to be five-category GRM items, although category collapsing occurred to a small proportion of replications. Given that a variety of IBA models has been introduced (Cai et al., 2011), future investigations could extend the longitudinal DIF study to include a different item type (e.g., the IBA extension of the three parameter logistic model) or a mixture of different types of items.

In addition, in the current study, the performance of IBA-SEM/XPD paired with Wald was not compared to other longitudinal DIF (or measurement invariance) methods found in recent literature (e.g., Coertjens, Donche, De Maeyer, Vanthournout, & Van Petegem, 2012; Mukherjee, Gibbons, Kristjansson, & Crane, 2013). Future research could focus on the comparisons of various approaches designed to detect longitudinal DIF (or establish measurement invariance), and find out which method is most appropriate under different simulation conditions.

Last but not least, the longitudinal DIF approach proposed by the current study, like many of the multiple-group DIF methods, requires an anchor item/subset for linking multiple sets of cross-time data on the same latent metric. Although there exist a variety of procedures designed to select group-invariant anchors in a multiple-group context (e.g., the rank-based anchor selection procedure; Woods, 2009), no research has shed light on the selection of time-invariant anchors for longitudinal DIF testing purposes. Therefore, developing a sound procedure for the identification of time-invariant anchors becomes an imminent task for researchers who plan to conduct longitudinal DIF analysis.

Appendix A

Figure 2. Densities of Latent Mean Estimates at Time-2.

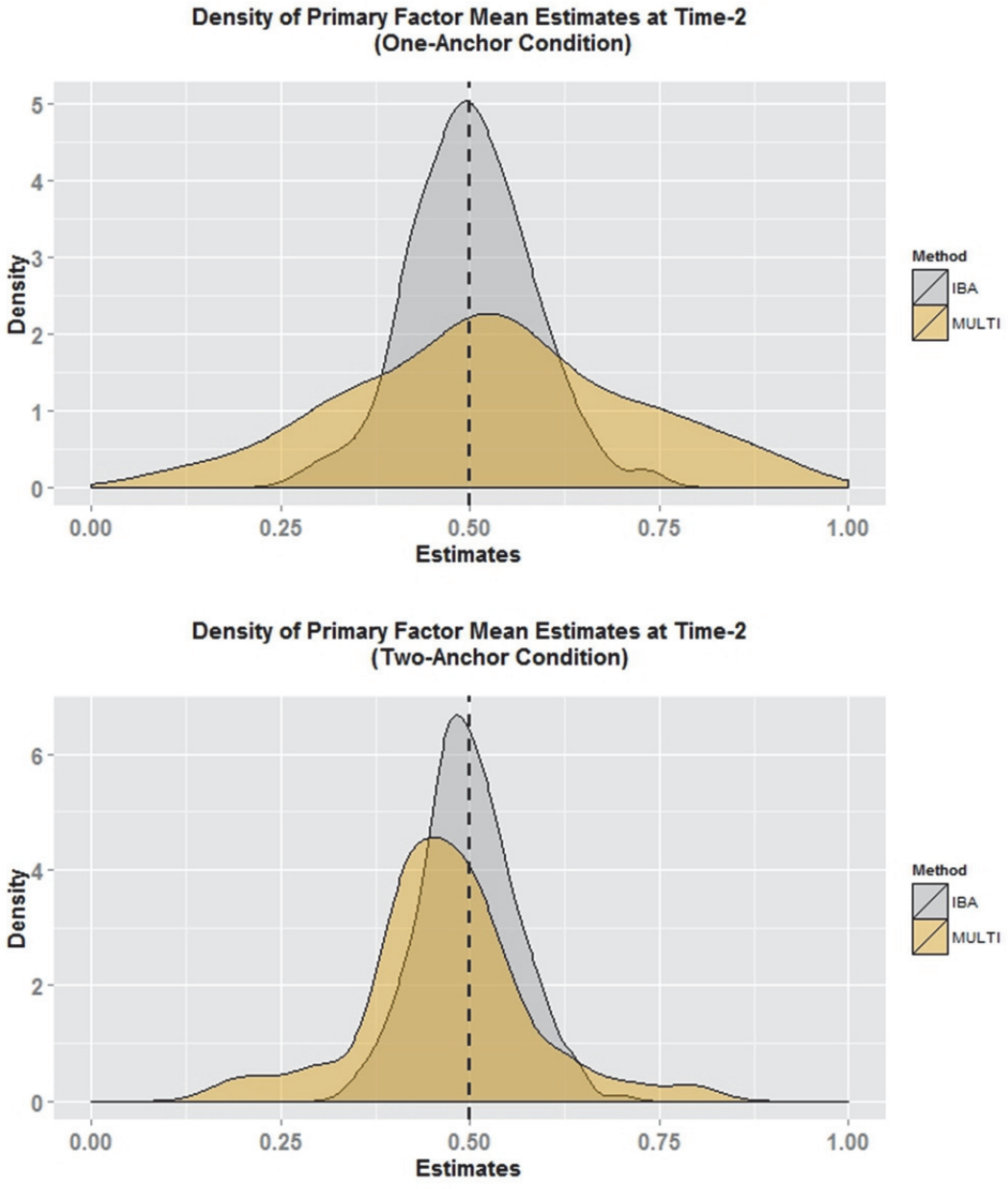


Figure 3. Densities of Latent Mean Estimates at Time-3.

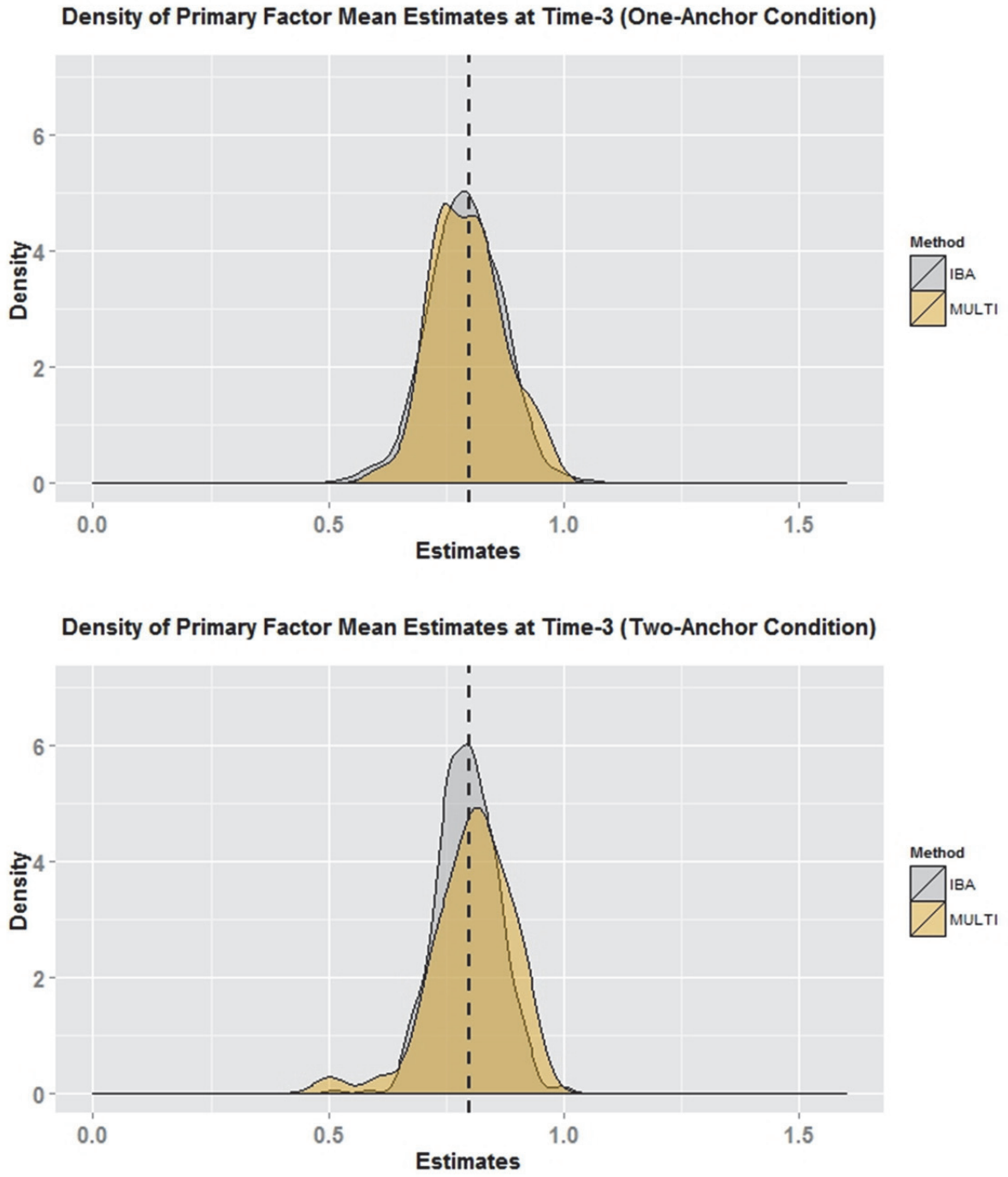


Figure 4. RMSEs of Latent Mean Estimation.

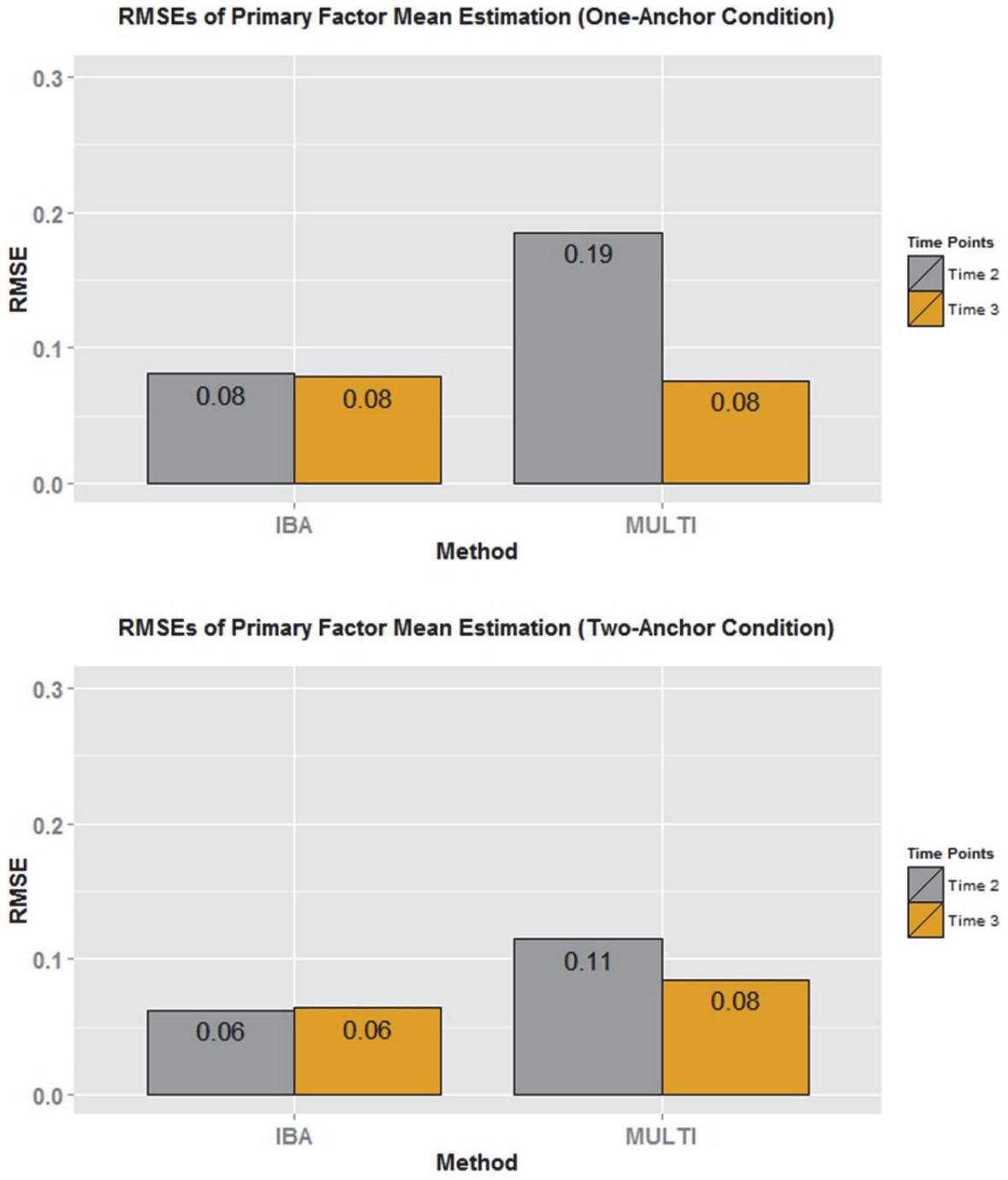


Figure 5. MAPEs of Latent Mean Estimation.

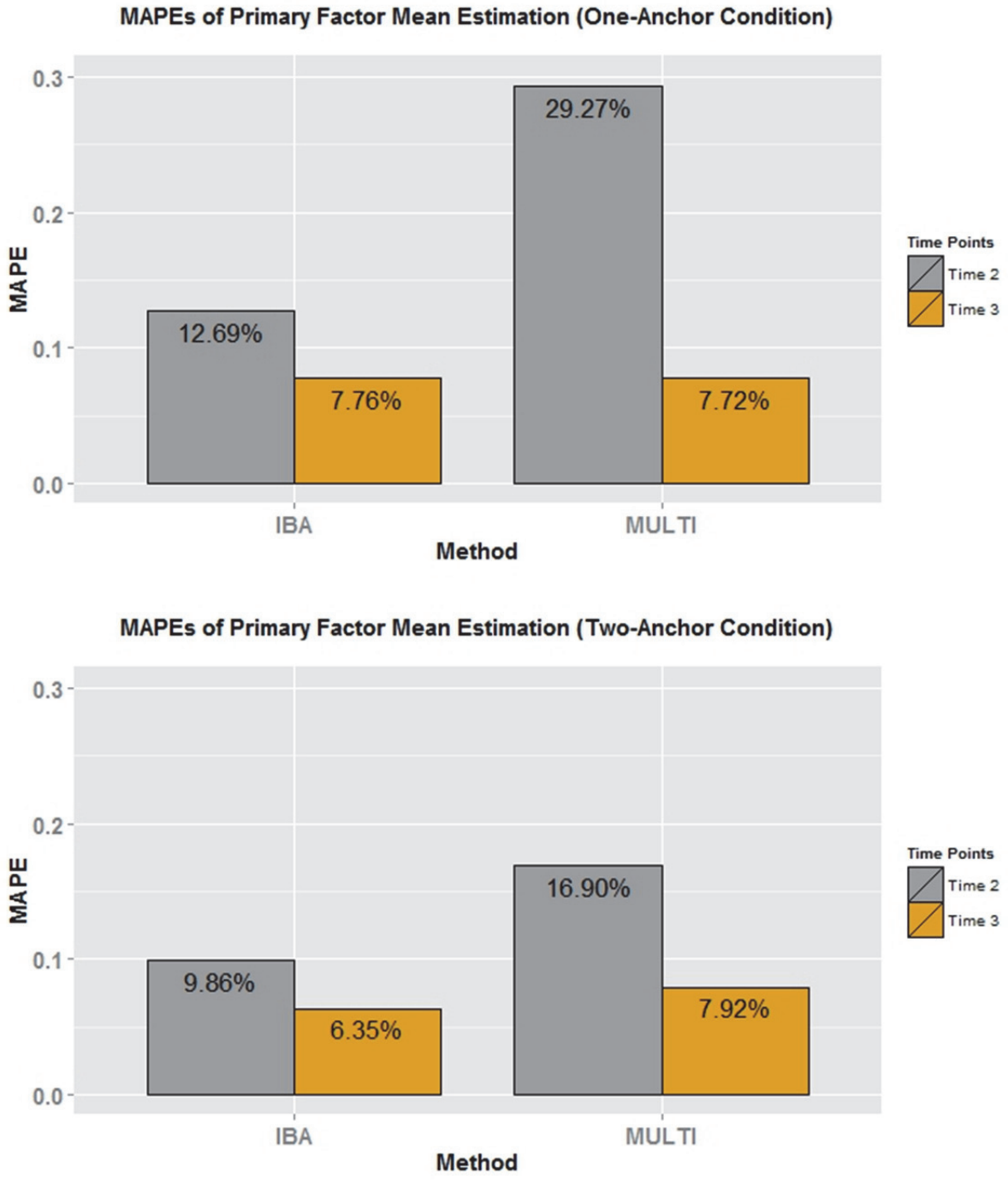


Figure 6. Densities of Latent SD Estimates at Time-2.

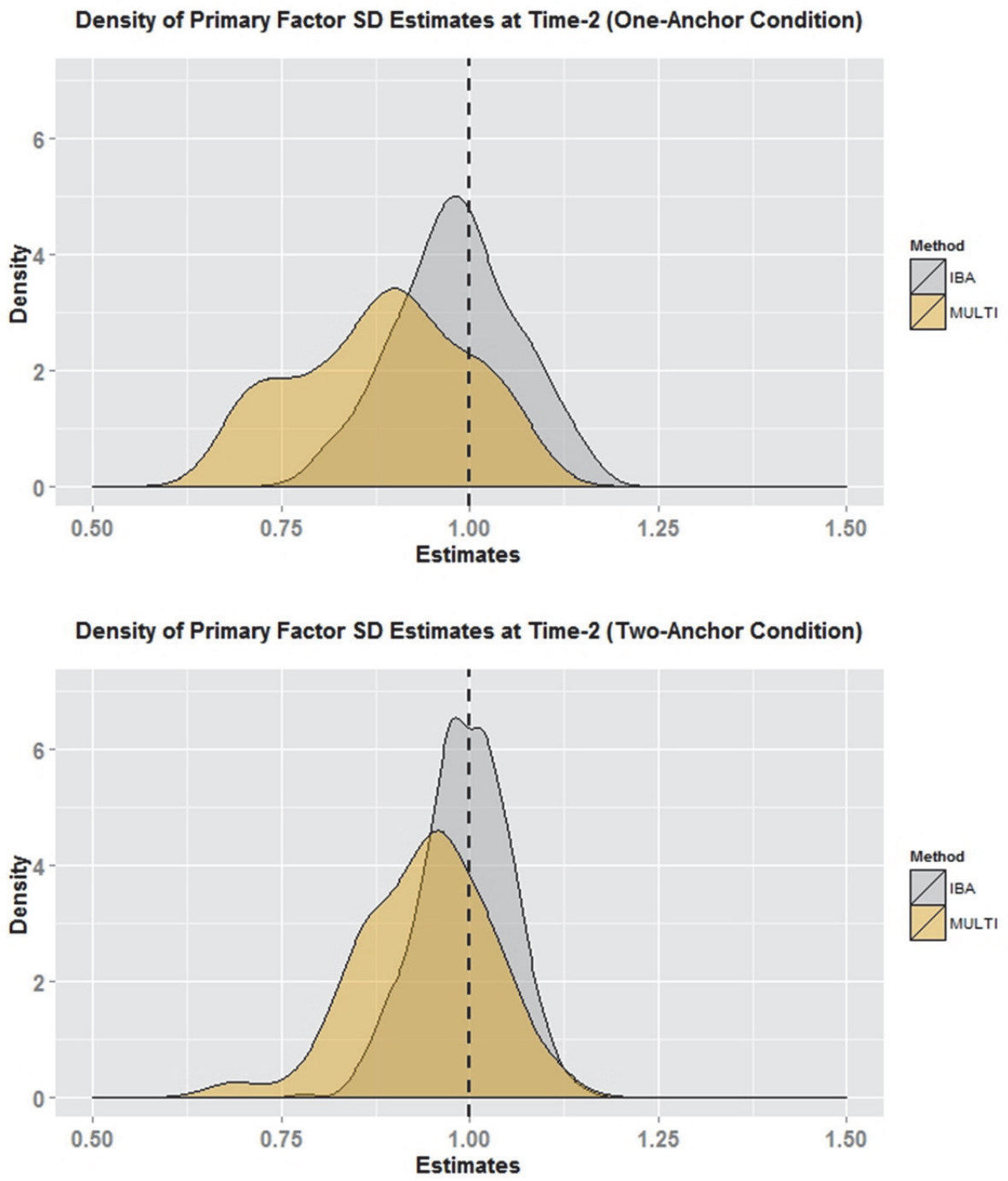


Figure 7. Densities of Latent SD Estimates at Time-3.

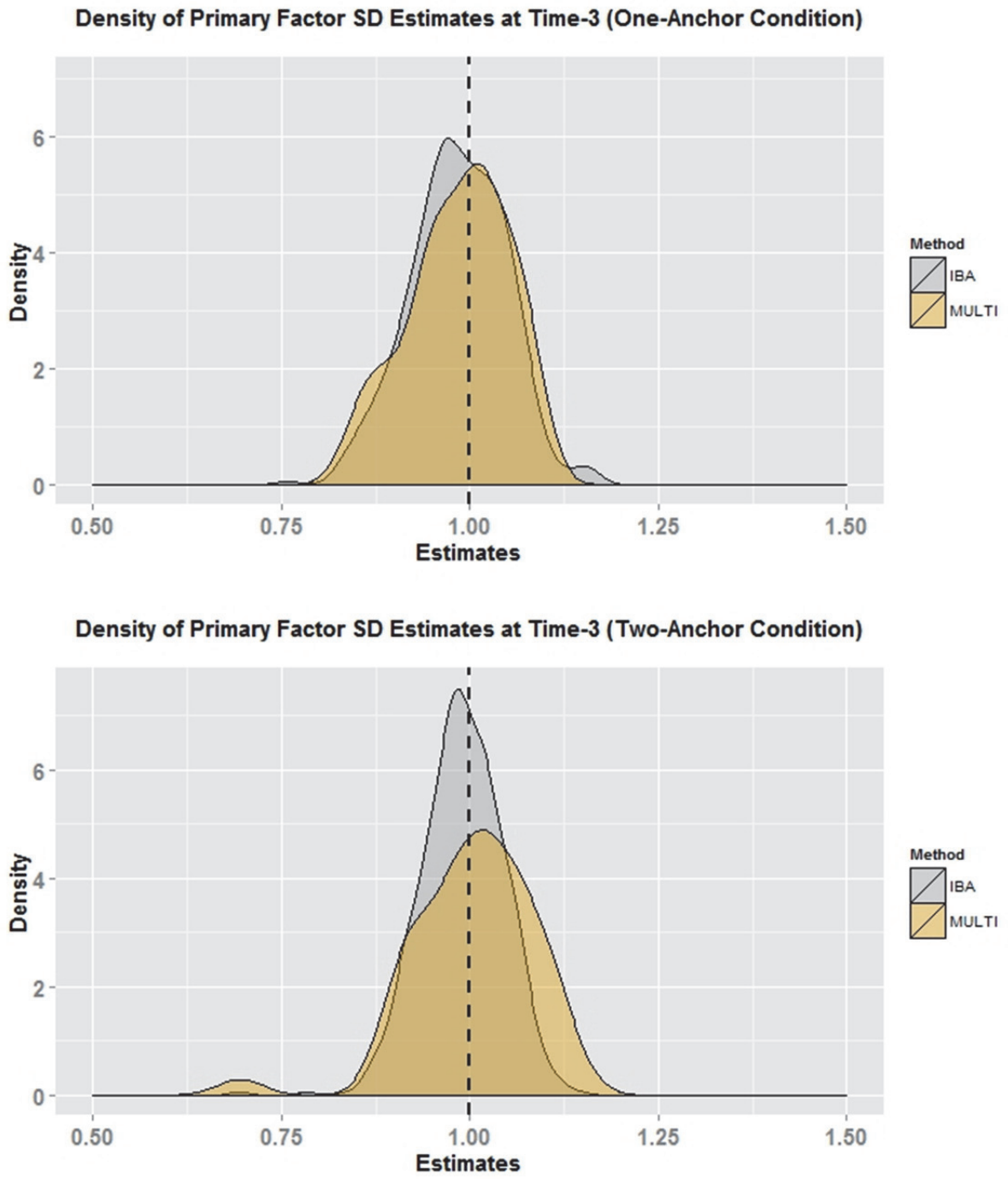


Figure 8. RMSEs of Latent SD Estimation.

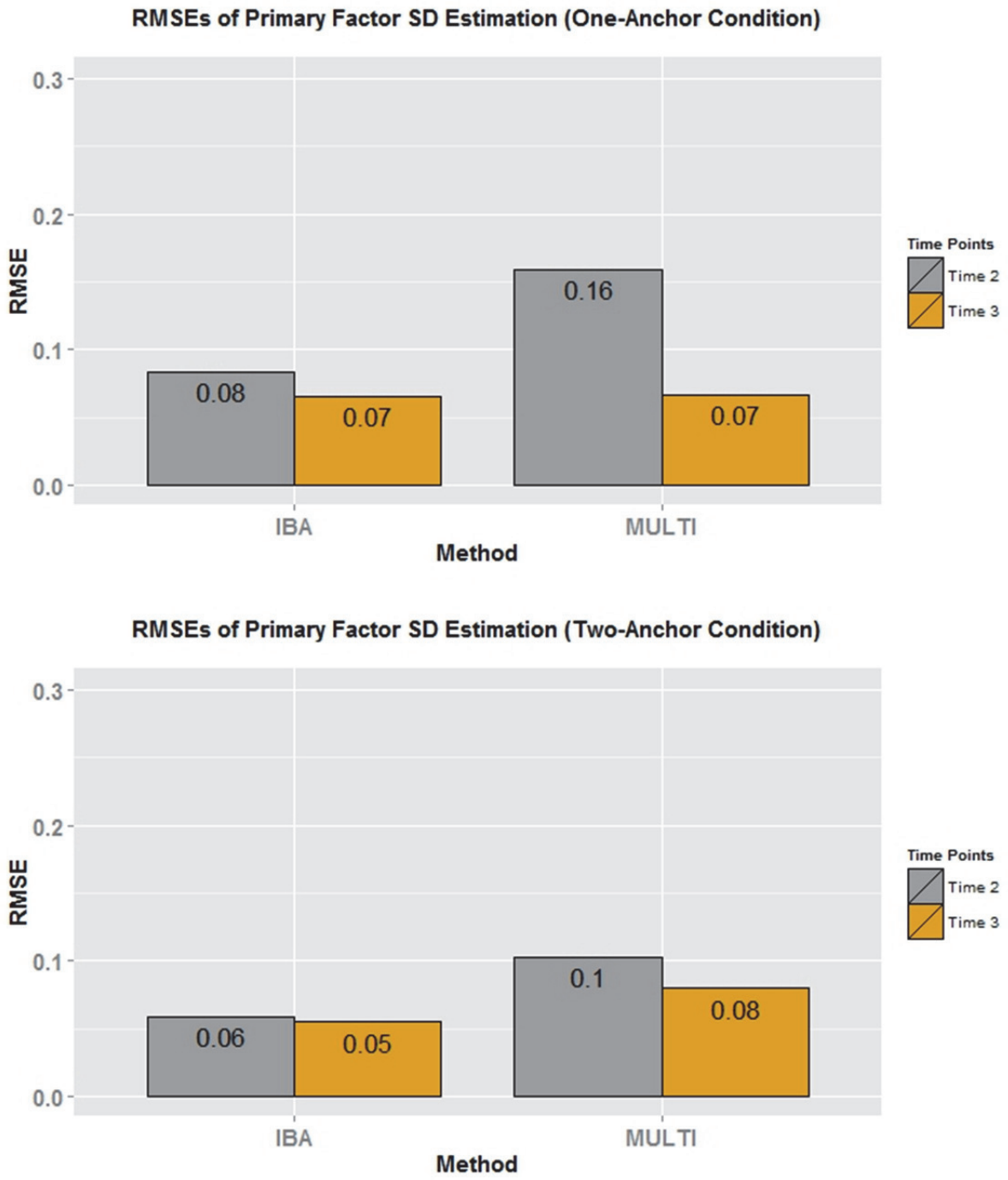


Figure 9. MAPEs of Latent SD Estimation.

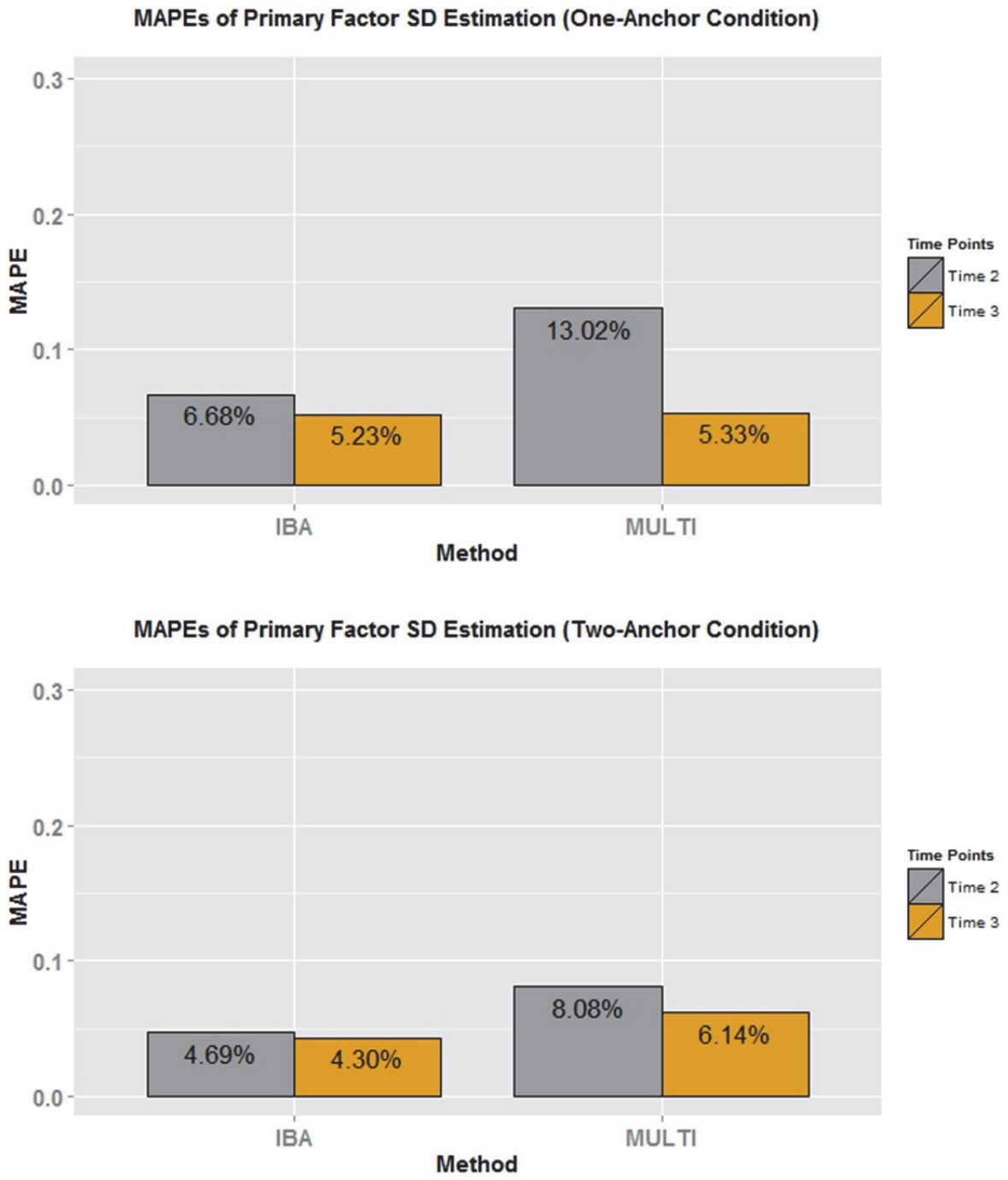


Figure 10. Densities of Latent Correlation Estimates (IBA only).

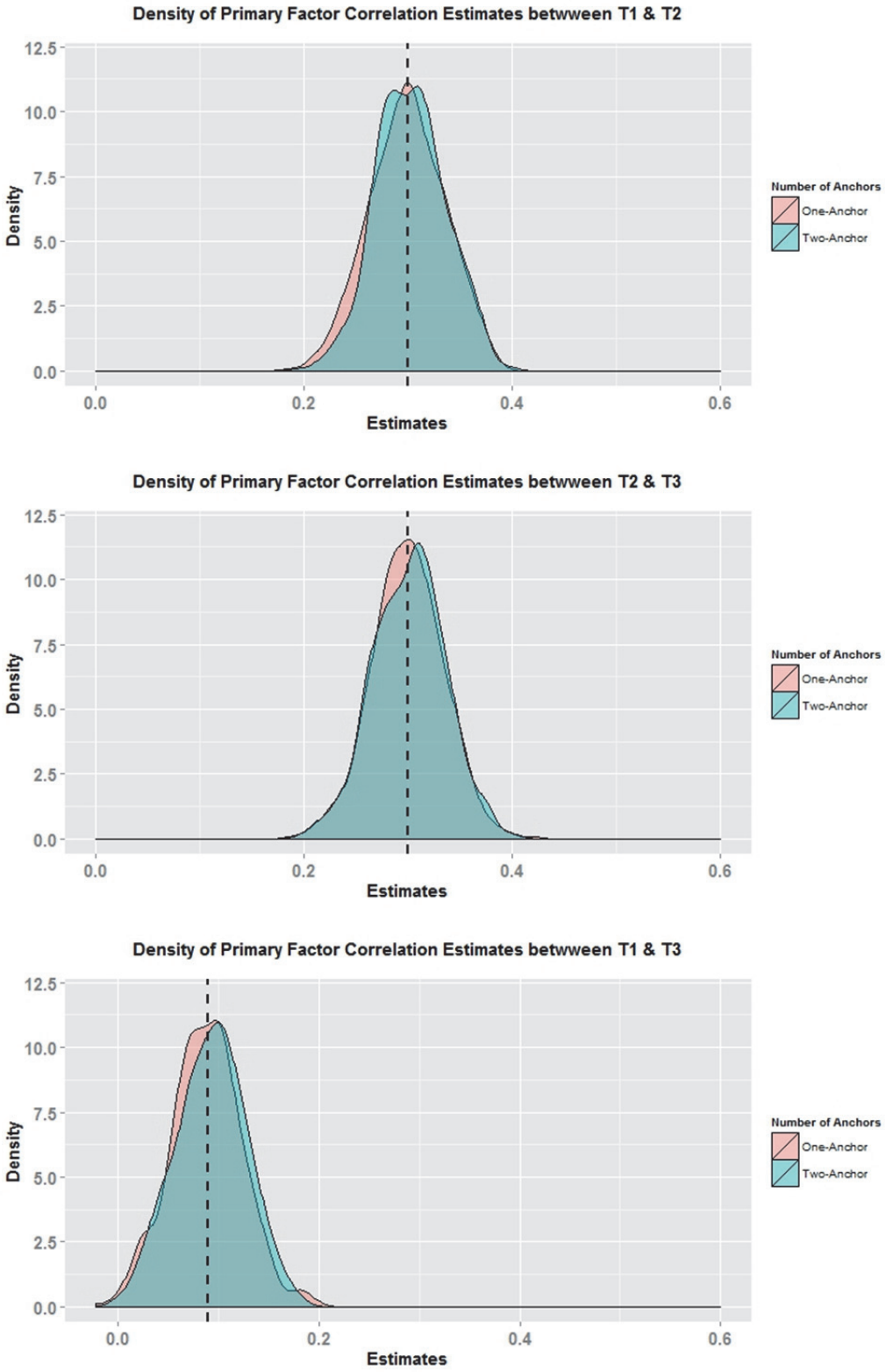


Figure 11. RMSEs and MAPEs of Latent Correlation Estimates (IBA only).

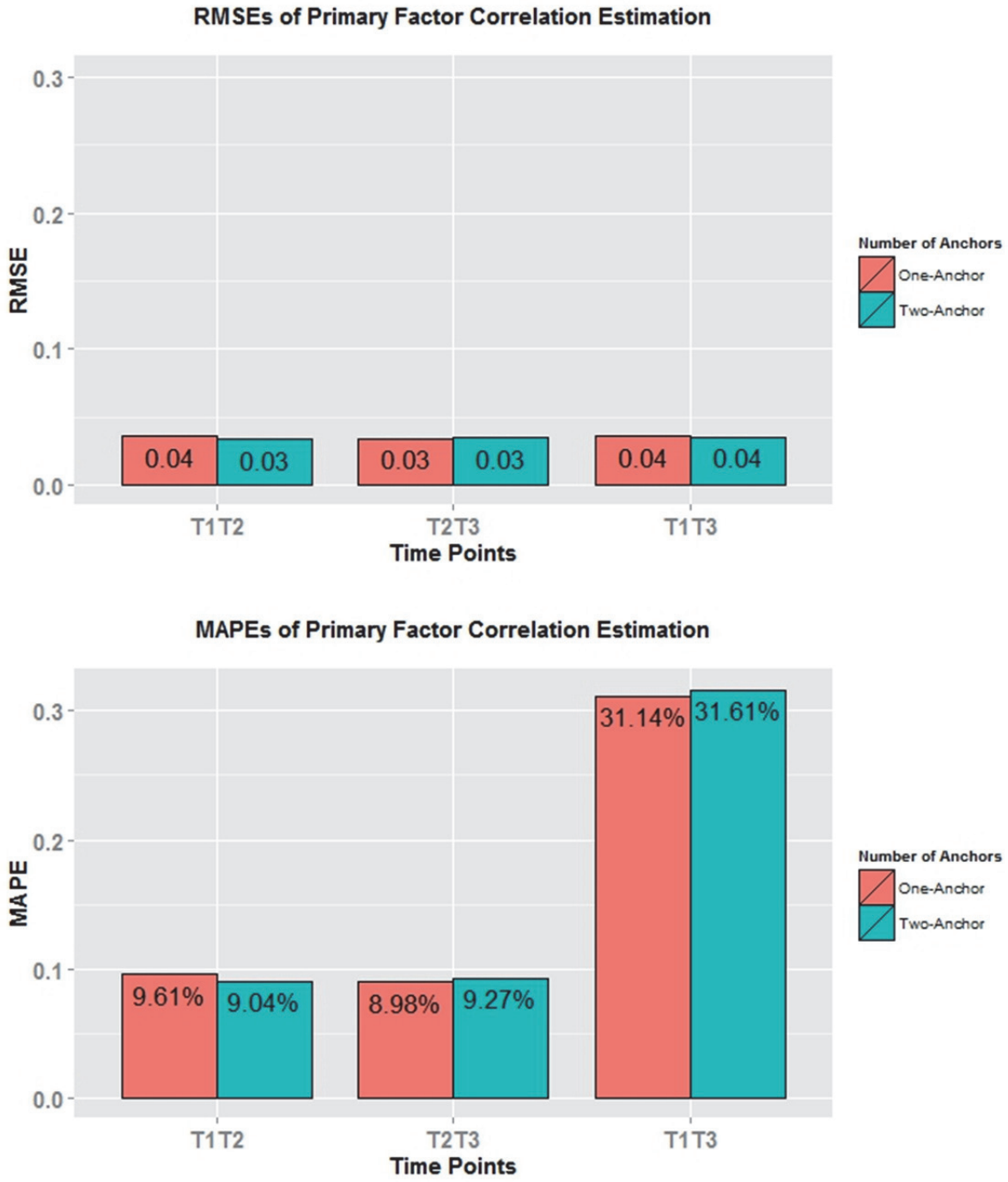
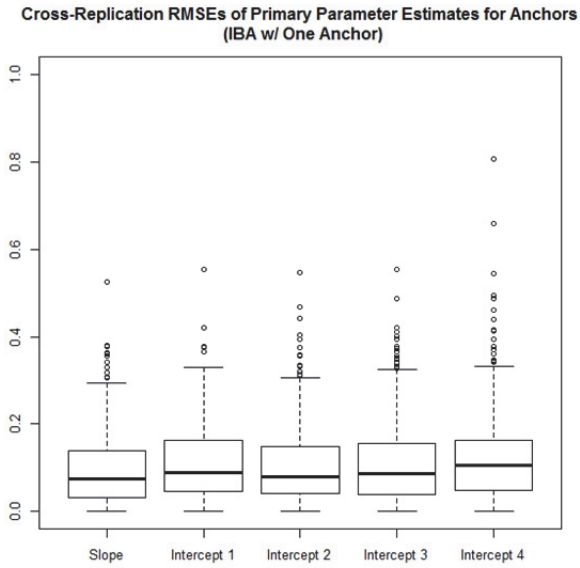
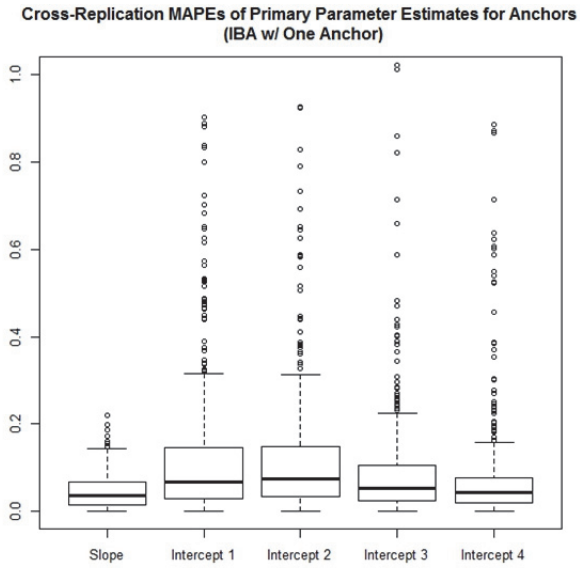
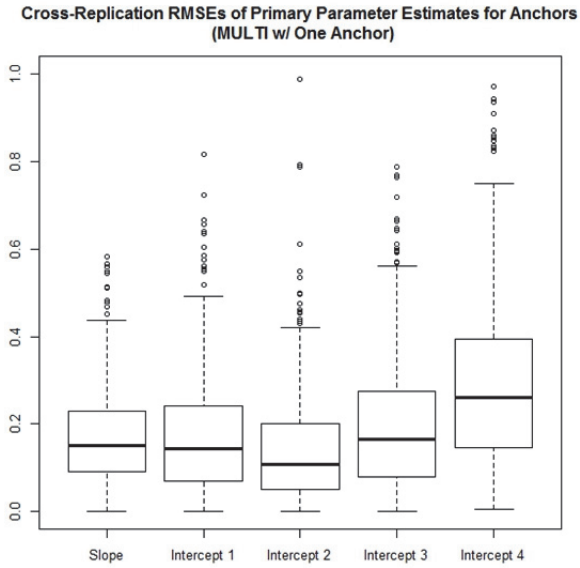


Figure 12. RMSEs and MAPEs of Primary Parameter Estimates for Anchors (One-Anchor).







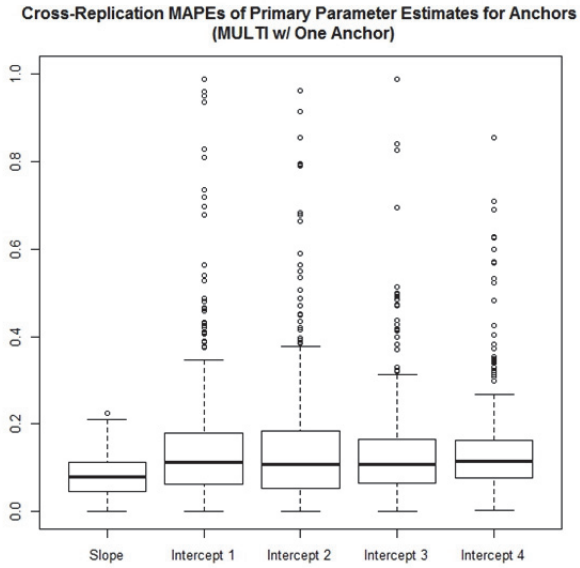


Figure 13. RMSEs and MAPEs of Primary Parameter Estimates for Anchors (Two-Anchor).

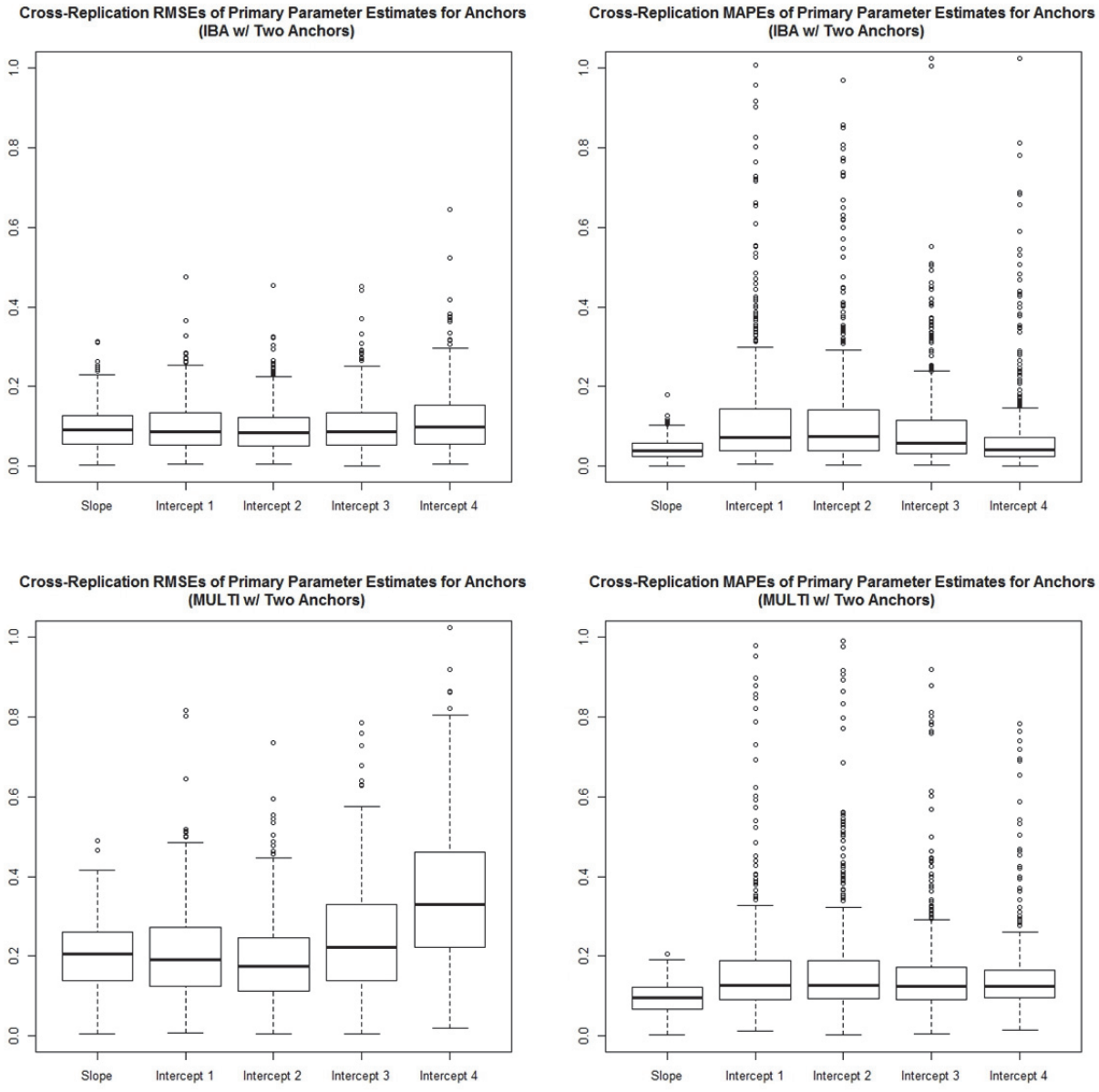


Figure 14. RMSEs and MAPEs of Primary Parameter Estimates for Time-1 Non-Anchors (One-Anchor).

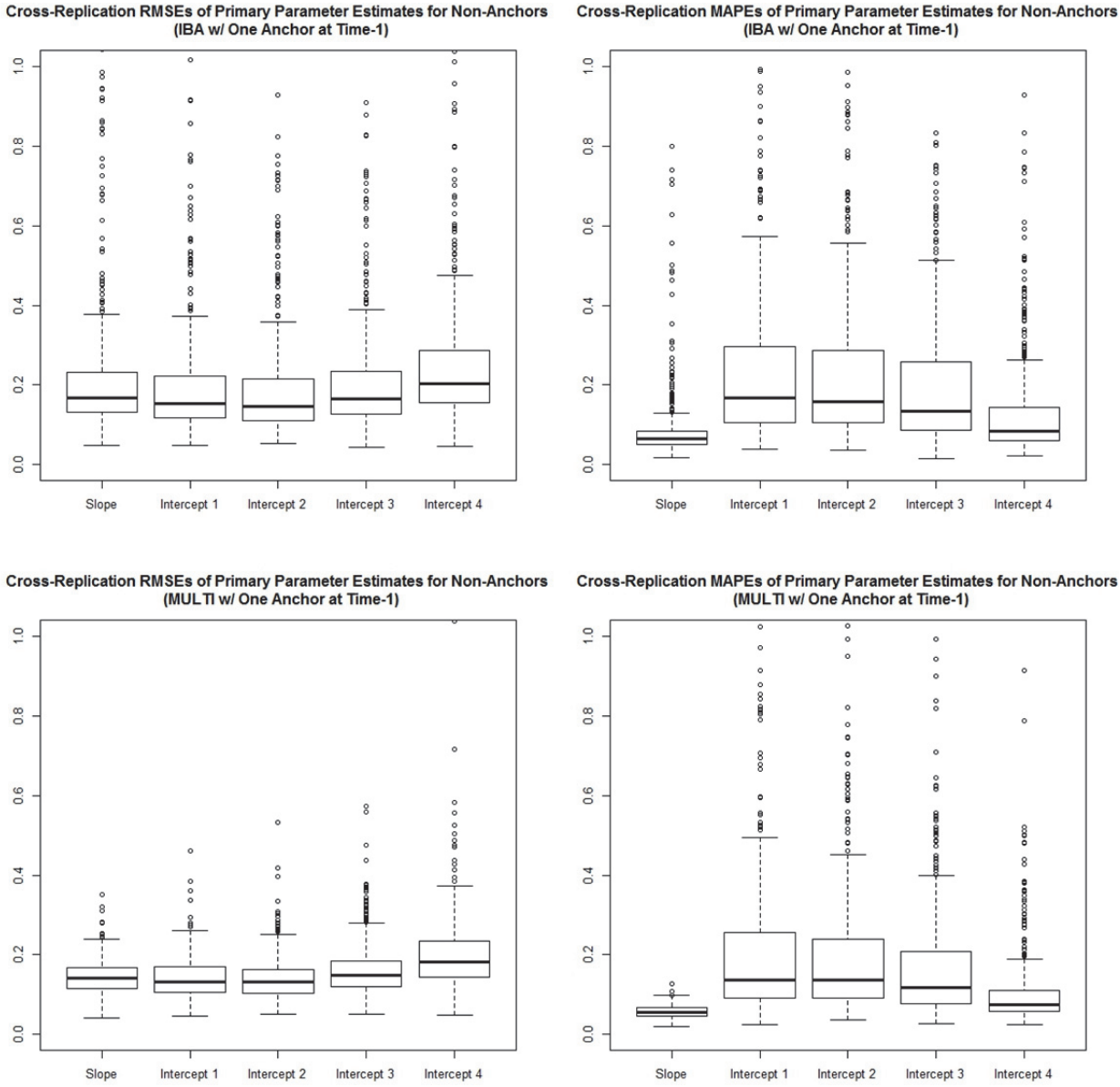


Figure 15. RMSEs and MAPEs of Primary Parameter Estimates for Time-1 Non-Anchors (Two-Anchor).

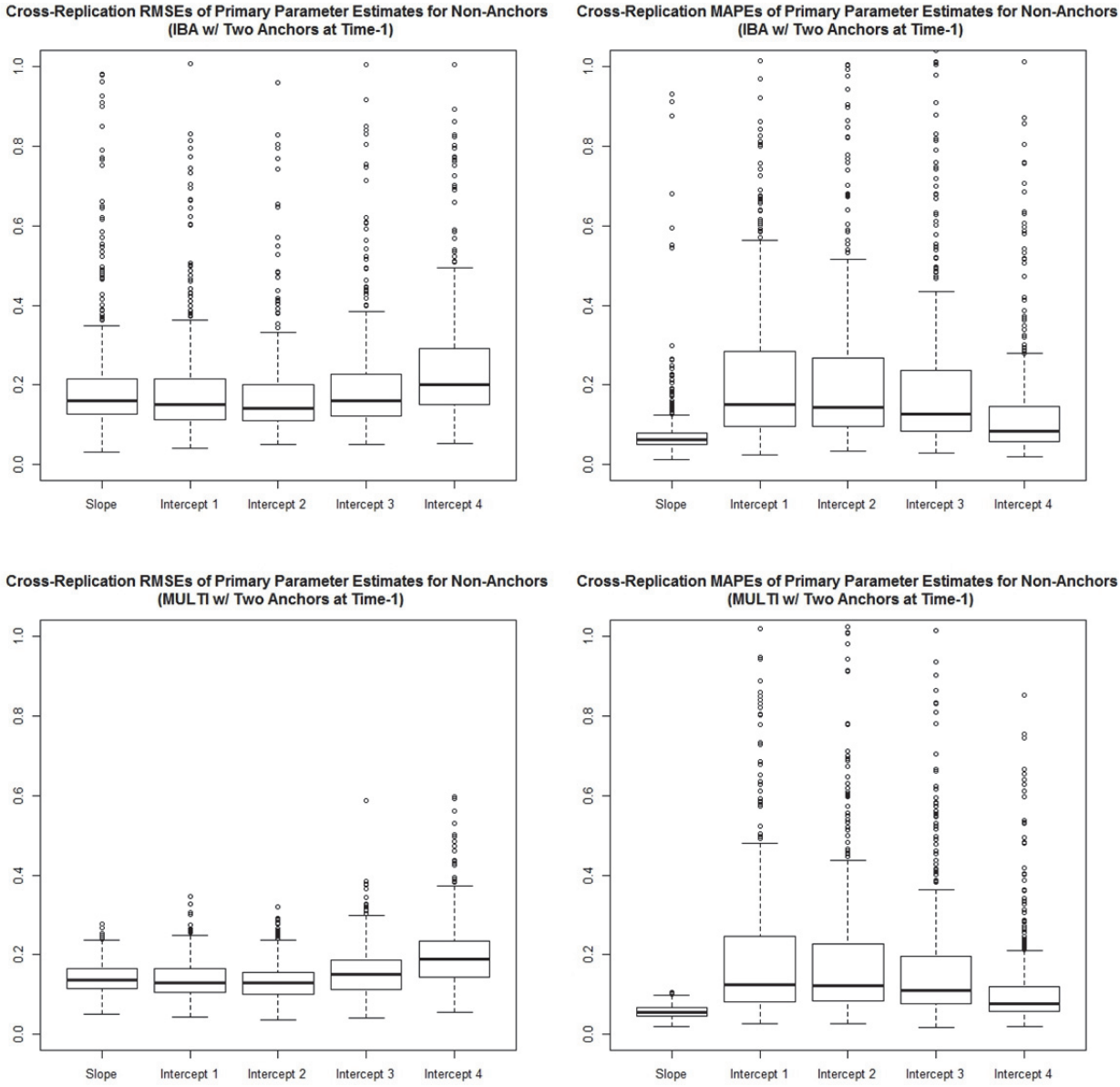


Figure 16. RMSEs and MAPEs of Primary Parameter Estimates for Time-2 Non-Anchors (One-Anchor).

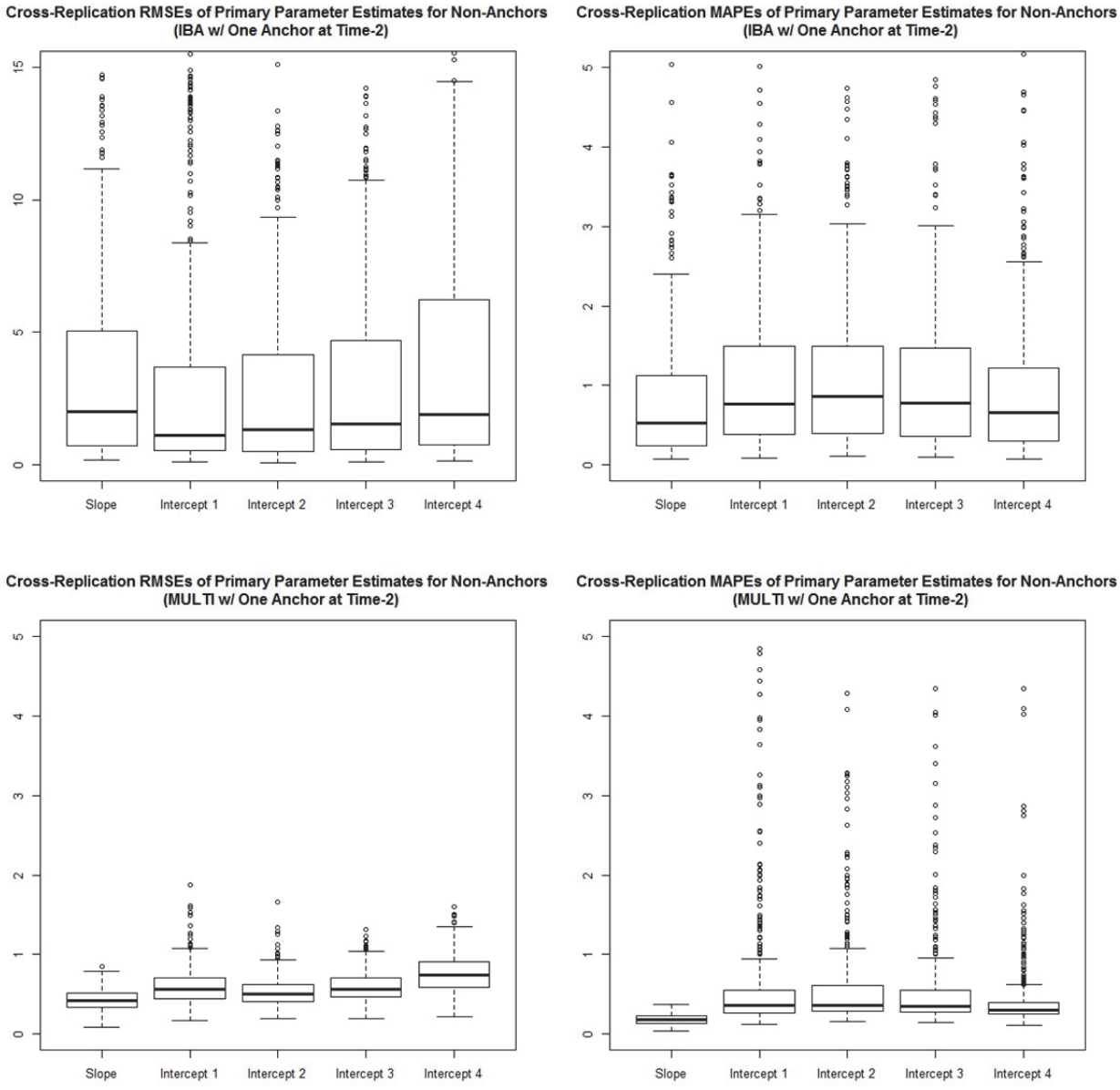


Figure 17. RMSEs and MAPEs of Primary Parameter Estimates for Time-2 Non-Anchors (Two-Anchor).

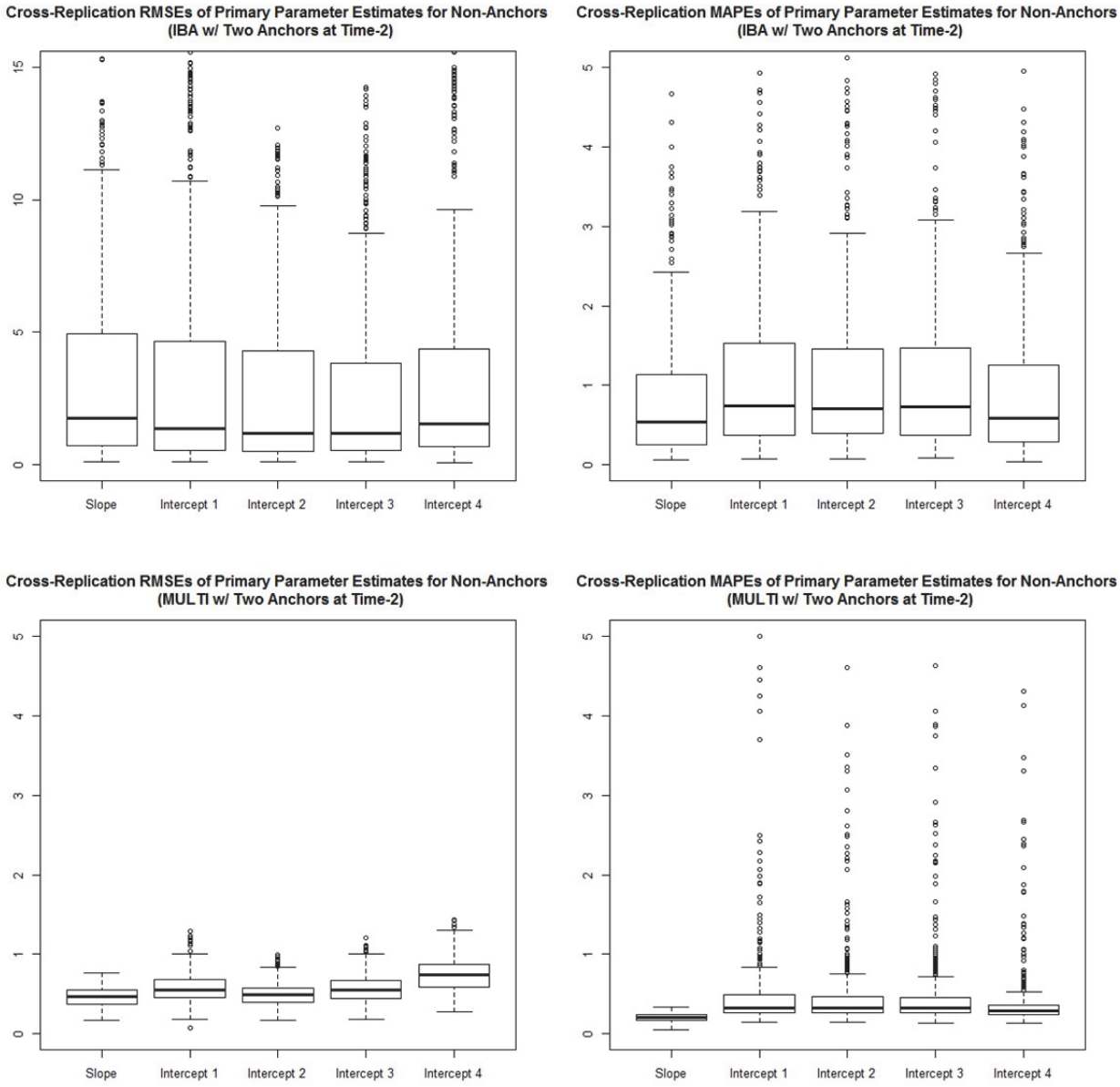


Figure 18. Absolute Percentage Deviations of All IBA-Estimated Primary Parameters for Time-2 Non-Anchors.

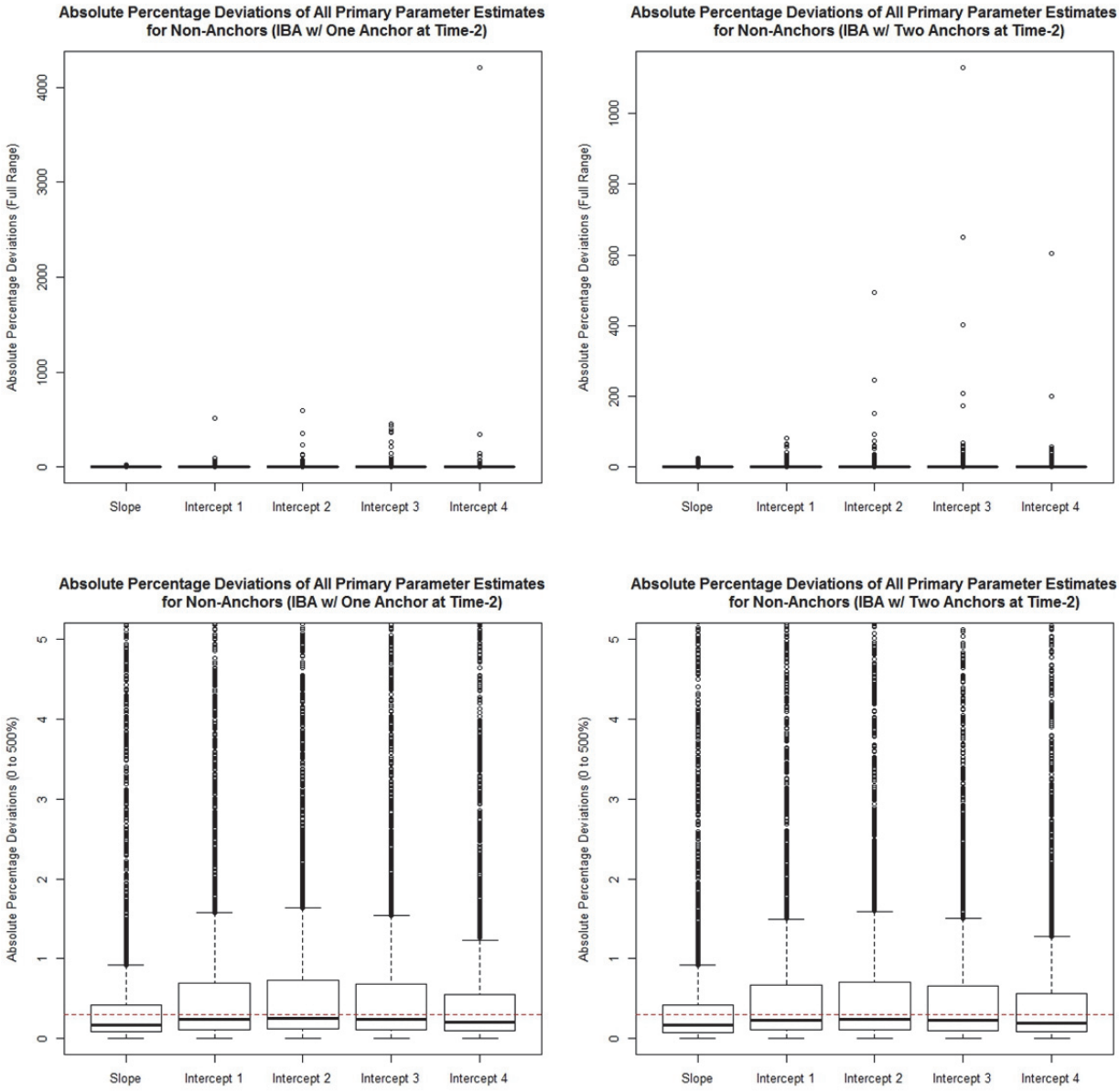


Figure 19. RMSEs and MAPEs of Primary Parameter Estimates for Time-3 Non-Anchors (One-Anchor).

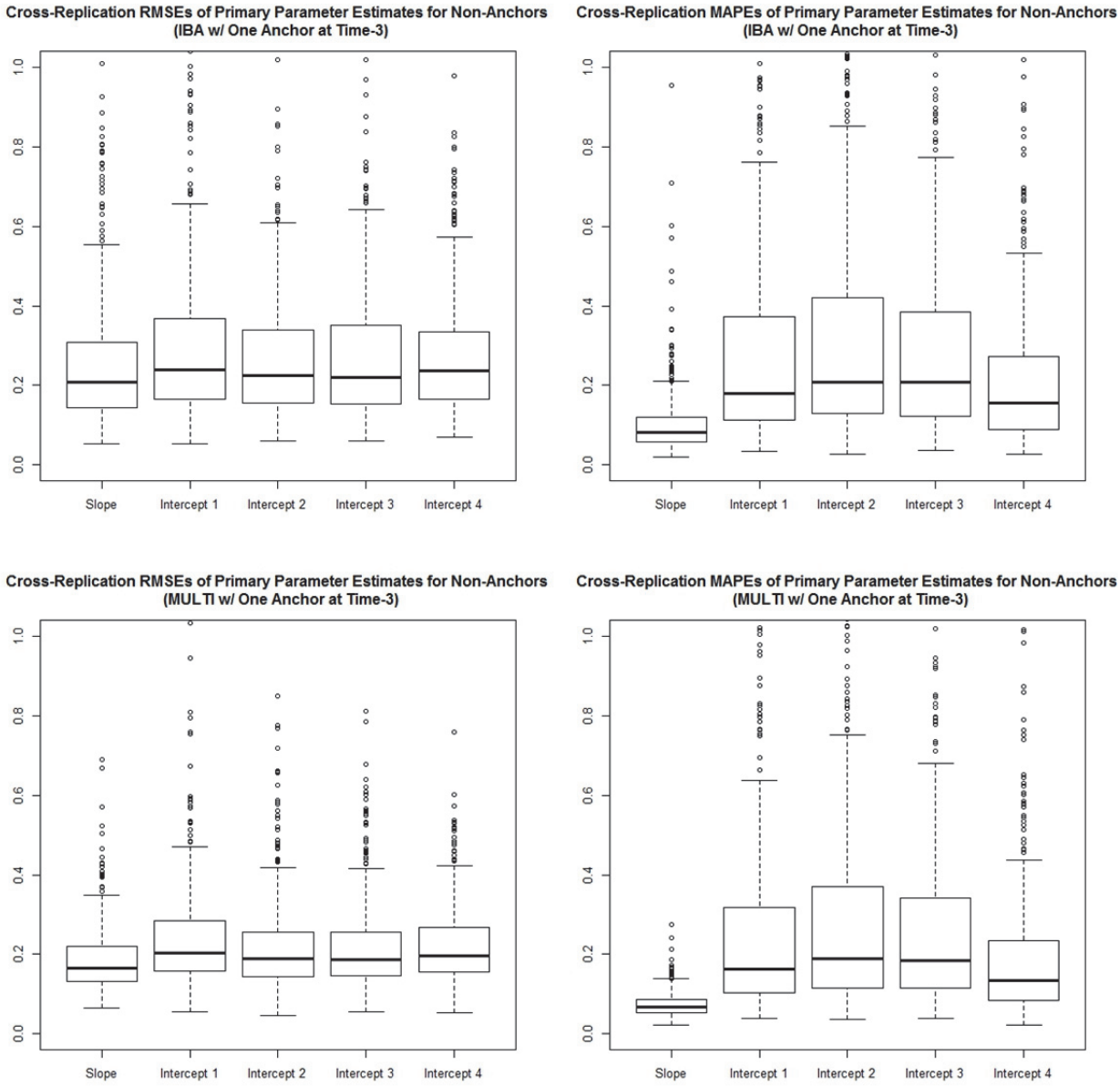


Figure 20. RMSEs and MAPEs of Primary Parameter Estimates for Time-3 Non-Anchors (Two-Anchor).

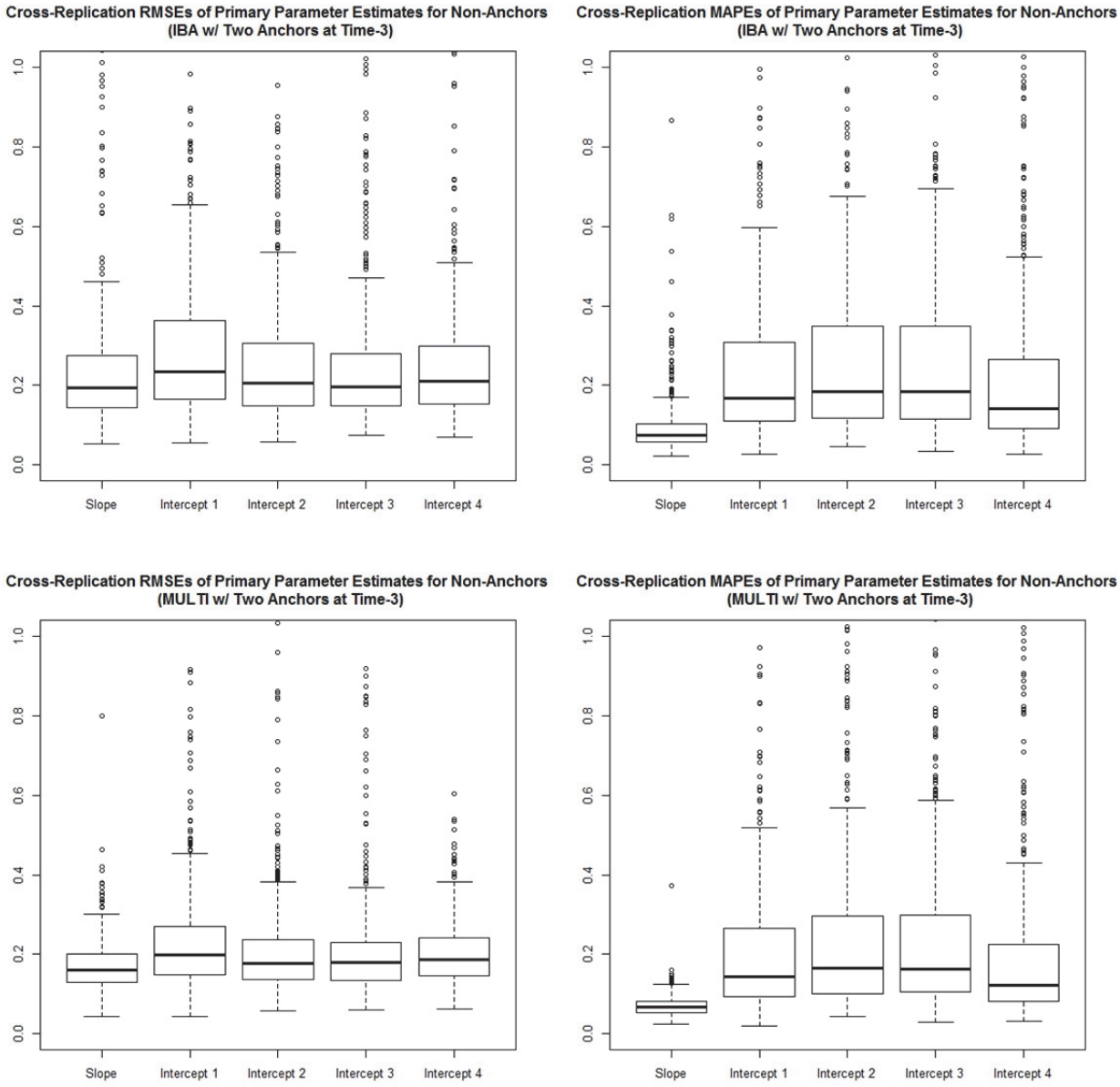
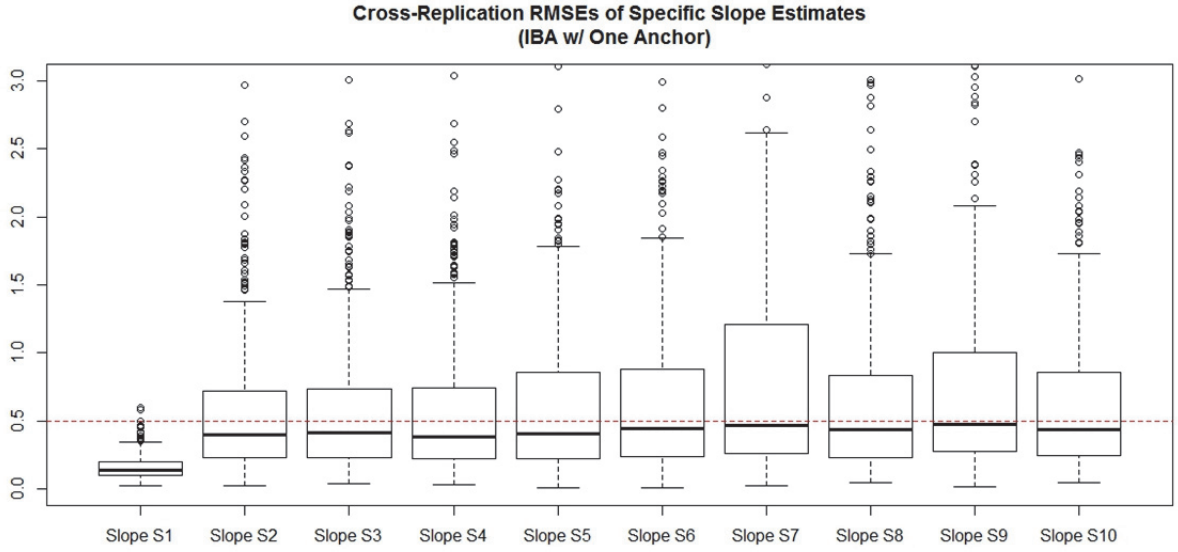


Figure 21. RMSEs of Specific Slope Estimates (for IBA Only).



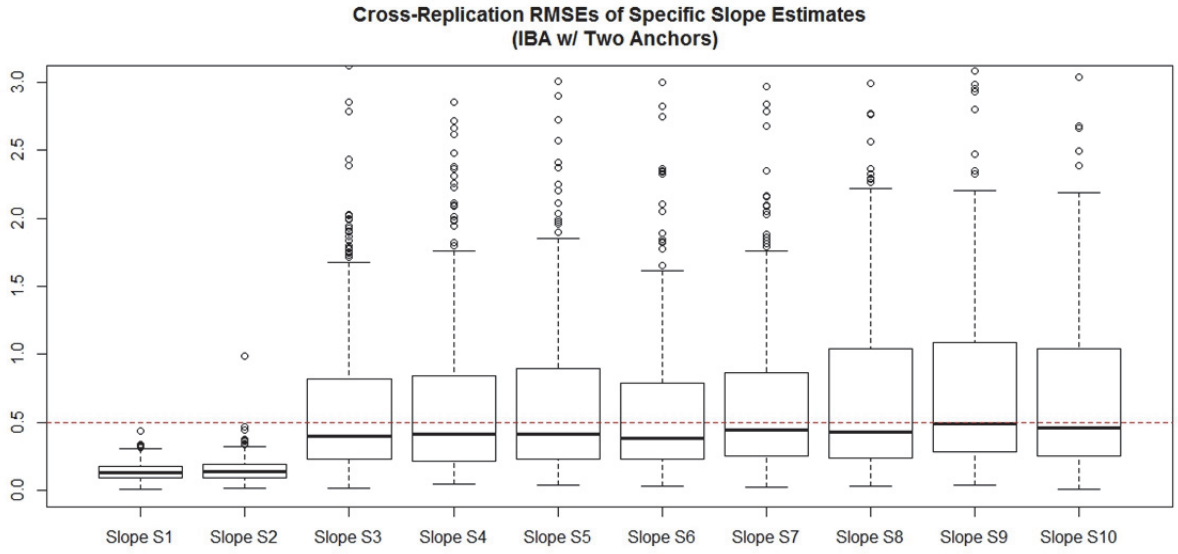


Figure 22. MAPEs of Specific Slope Estimates (for IBA Only).

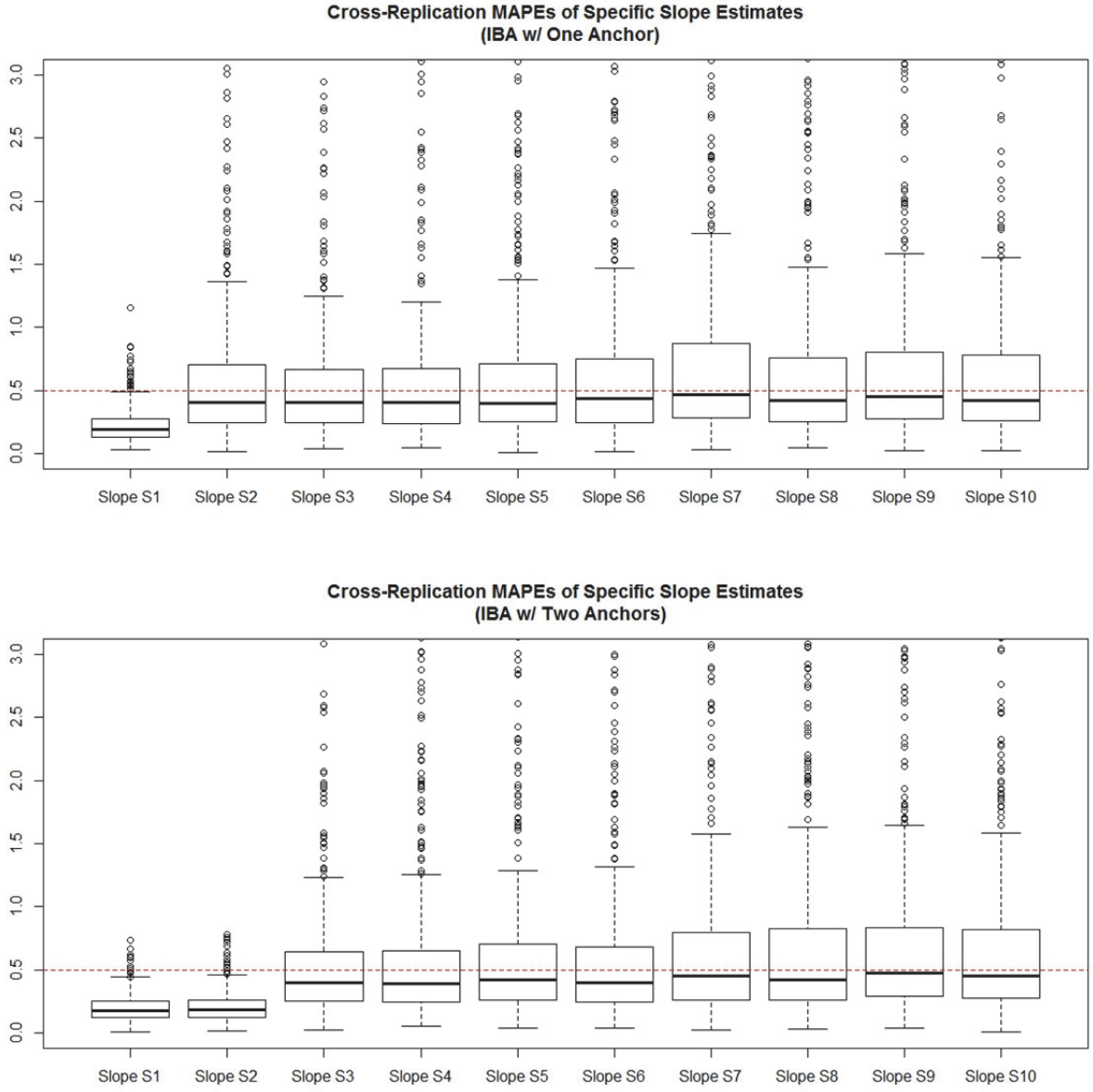


Figure 23. Statistical Power and Type I Error of Omnibus DIF Detection.

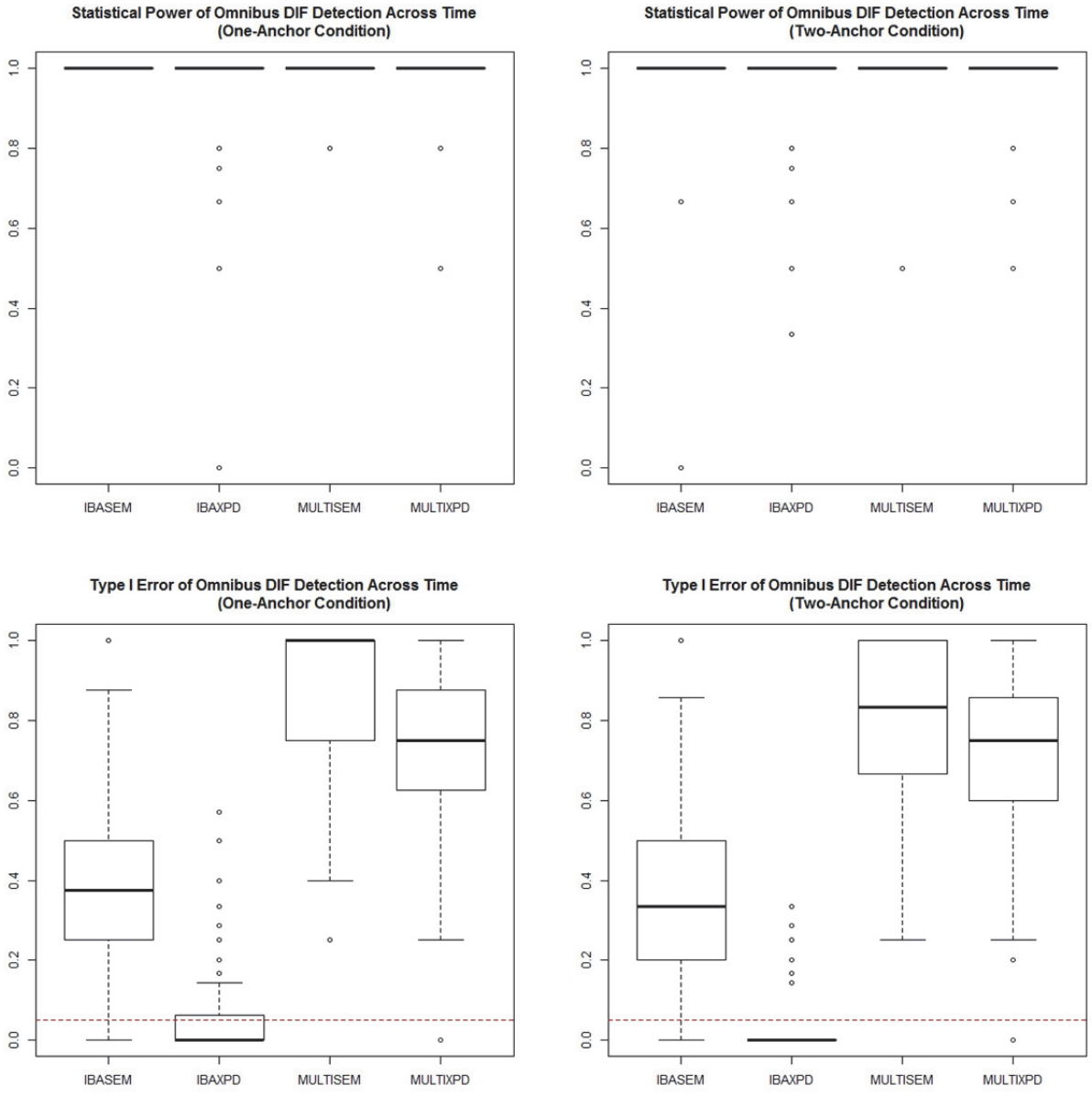


Figure 24. Statistical Power and Type I Error of DIF Detection Comparing Time-1 and Time-2.

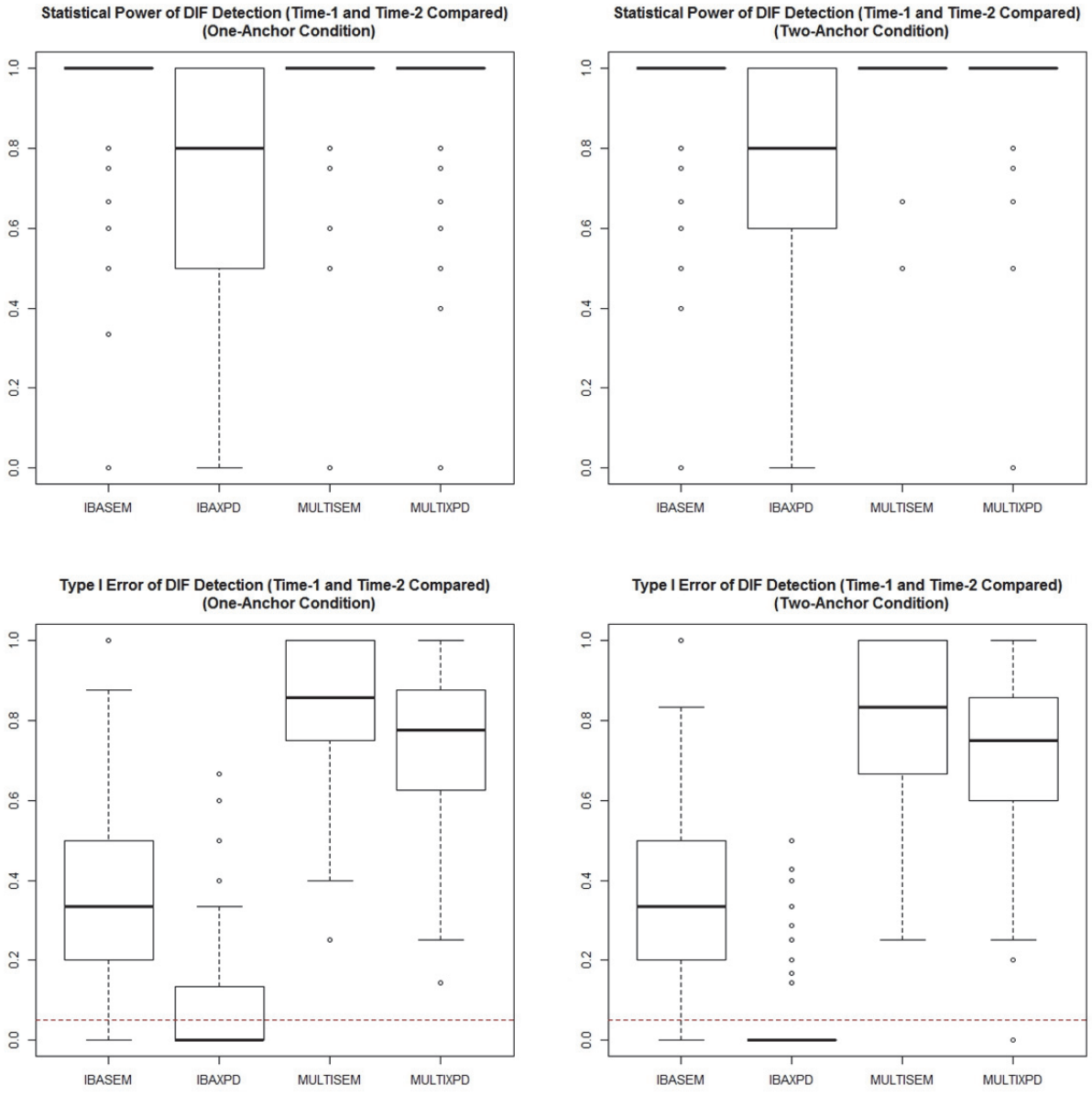


Figure 25. Statistical Power and Type I Error of DIF Detection Comparing Time-1 and Time-3.

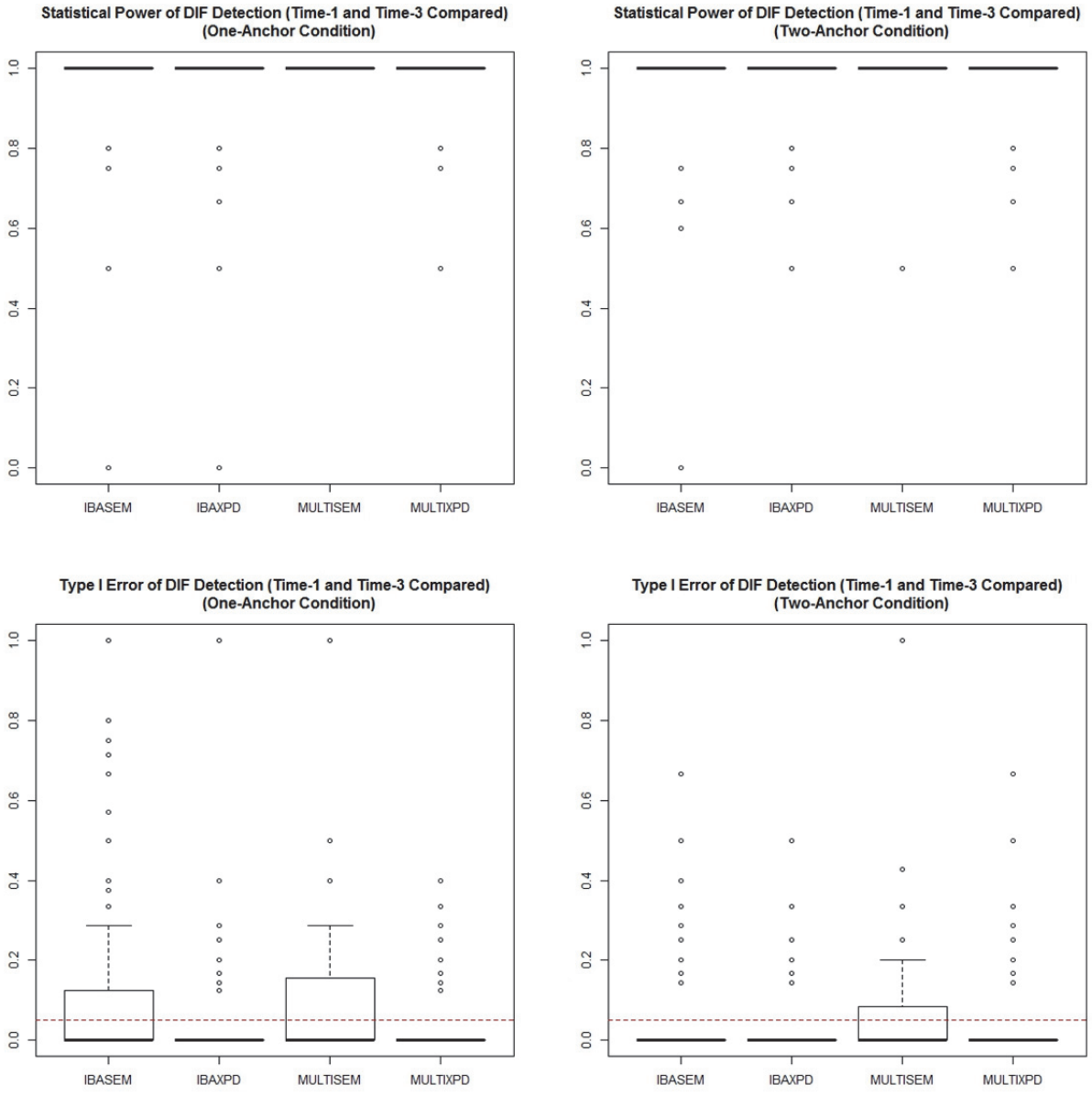
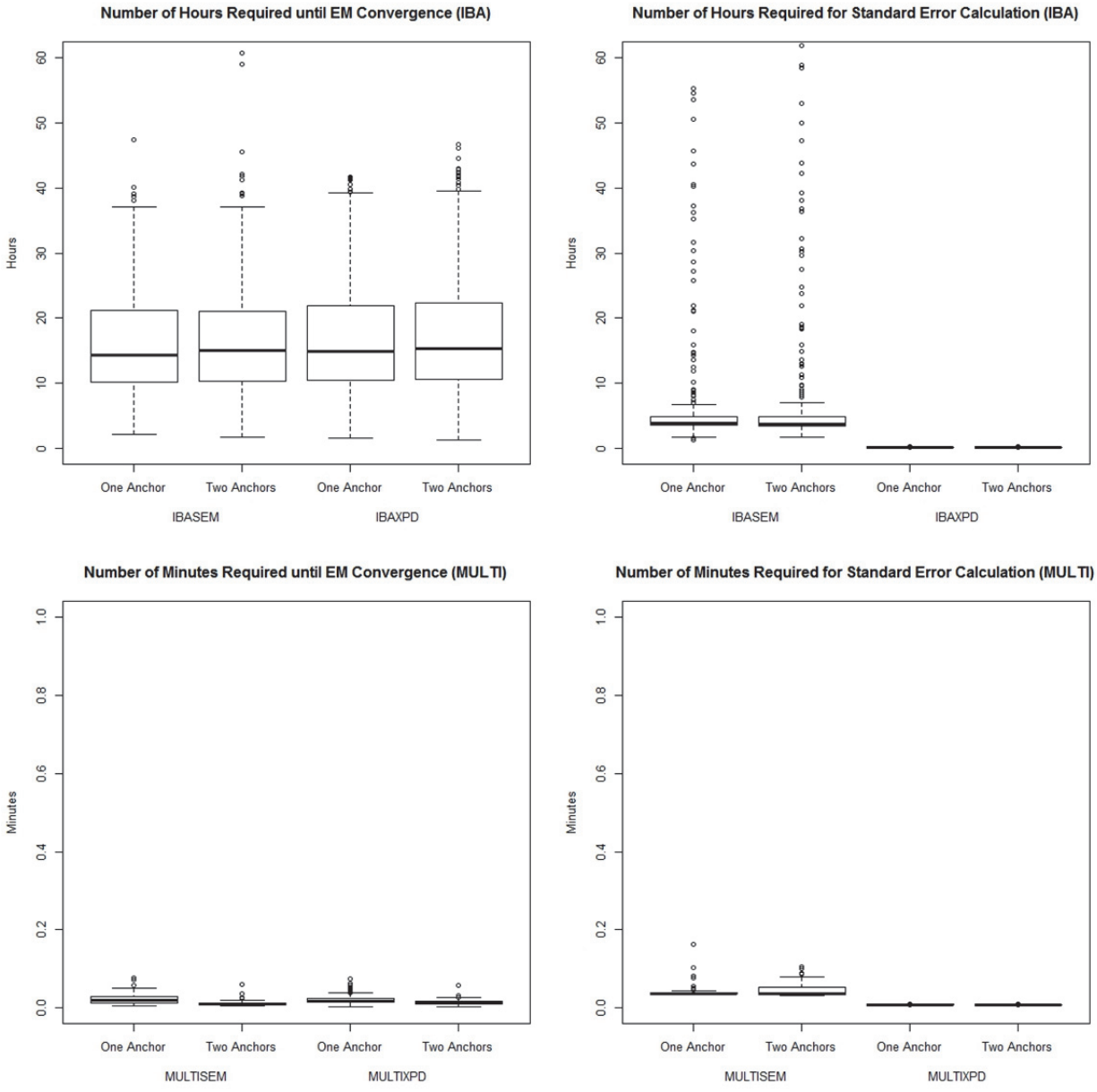


Figure 26. Computation Time Required for Model Estimation and Standard Error Calculation.



Appendix B

Table 1. Average Type I Errors of Omnibus DIF Tests.

	IBA		MULTI	
	SEM	**XPD**	**SEM**	**XPD**
One Anchor	.374 (.221)	.048 (.094)	.872 (.170)	.758 (.184)
Two Anchors	.347 (.230)	.035 (.083)	.793 (.207)	.733 (.207)

* Standard deviations associated with the average Type I errors were listed inside parentheses.

Table 2. Average Type I Errors of Pairwise DIF Tests between Time-1 and Time-2.

	IBA		MULTI	
	SEM	**XPD**	**SEM**	**XPD**
One Anchor	.356 (.212)	.063 (.108)	.820 (.196)	.763 (.186)
Two Anchors	.343 (.229)	.046 (.094)	.788 (.209)	.743 (.204)

* Standard deviations associated with the average Type I errors were listed inside parentheses.

Table 3. Average Type I Errors of Pairwise DIF Tests between Time-1 and Time-3.

	IBA		MULTI	
	SEM	**XPD**	**SEM**	**XPD**
One Anchor	.067 (.137)	.001 (.059)	.090 (.173)	.023 (.063)
Two Anchors	.041 (.094)	.001 (.048)	.078 (.180)	.021 (.070)

* Standard deviations associated with the average Type I errors were listed inside parentheses.

Table 4. Primary Item Parameter Estimates for the 10-Item MFQ Subscale (Two Waves).

	Slope			**Intercepts**			
Year 1994	a	d_1	d_2	d_3	d_4	d_5	d_6
v1	0.87	5.12	3.65	2.22	-0.10	-1.55	-3.63
v2	0.63	3.75	1.73	0.39	-2.18	-3.80	-5.51
v3	0.90	6.18	3.95	2.67	0.75	-0.38	-2.37
v4	0.86	4.87	3.83	1.89	0.09	-1.08	-3.26
v5	0.96	5.72	4.18	2.74	1.30	0.01	-1.94
v6	0.93	5.23	3.95	2.20	0.43	-0.73	-2.55
v7	1.06	6.16	4.91	3.39	1.91	0.70	-1.47
v8	0.92	6.87	5.11	3.76	2.26	1.22	-0.42
v9	7.11	12.55	7.77	3.02	-1.11	-5.83	-10.09
v10	7.25	11.02	5.56	1.86	-2.55	-6.46	-10.78
	Slope			**Intercepts**			
Year 1997	a	d_1	d_2	d_3	d_4	d_5	d_6
v1	0.87	5.12	3.65	2.22	-0.10	-1.55	-3.63
v2	0.72	3.59	1.68	0.23	-2.48	-4.63	-6.47
v3	0.76	6.14	3.87	2.67	0.78	-0.45	-2.59
v4	0.74	5.14	3.18	1.94	0.30	-1.02	-3.38
v5	0.82	4.87	3.30	2.43	0.91	-0.08	-2.30
v6	0.63	5.55	3.66	1.75	0.22	-0.96	-2.84
v7	0.60	7.25	4.49	2.98	1.24	0.05	-1.91
v8	0.62	6.11	4.80	3.52	1.86	0.99	-1.18
v9	7.90	12.82	8.34	3.50	-2.04	-6.26	-12.16
v10	7.10	9.74	5.45	1.65	-3.26	-6.90	-10.99

* Note: Item **v1** was designated as the time-invariant anchor, and all model parameters were estimated under the bifactor structure.

Table 5. IBA-XPD Results for Testing the 10-Item MFQ Subscale across Two Waves.

	Q (χ^2 $distributed$)	df	p
v2	1.81	7	0.970
v3	0.82	7	0.997
v4	4.87	7	0.675
v5	4.39	7	0.734
v6	2.75	7	0.907
v7	2.89	7	0.895
v8	3.06	7	0.879
v9	2.52	7	0.926
v10	1.80	7	0.970

* Note: Item v1 was designated as the time-invariant anchor.

Table 6. IBA-SEM Results for Testing the 10-Item MFQ Subscale across Two Waves.

	Q $(\chi^2\ distributed)$	df	p
v2	5.37	7	0.614
v3	1.42	7	0.985
v4	9.32	7	0.231
v5	8.36	7	0.302
v6	6.19	7	0.517
v7	12.10	7	0.097
v8	11.02	7	0.138
v9	6.29	7	0.507
v10	4.79	7	0.685

* Note: Item **v1** was designated as the time-invariant anchor.

Appendix C

Sample flexMIRT script used for modeling the 10-Item MFQ subscale across two waves:

```
<Project>
Title = "IBA-SEM/XPD Longitudinal Modeling Illustration";
Description = "IBA-SEM/XPD Longitudinal Modeling Illustration";

<Options>
Mode = Calibration;
//Number of quadrature points and the range (symmetric);
Quadrature = 15, 4.0;
//Save the covariance matrix as a separate file (REQUIRED);
SaveCOV = YES;
//Save the model parameters as a separate file (REQUIRED);
SavePRM = YES;
//Maximum number of EM cycles allowed;
MaxE = 5000;
//E-step convergence criterion;
Etol = 1e-3;
//M-step convergence criterion;
Mtol = 1e-5;
//Enable the 'SE' option below if using SEM procedure;
//SE = SEM;
//Enable the 'SEMtol' option below if adjusting SEM convergence criterion to be 0.005;
//SEMtol = 5e-3;

<Groups>
//Group name. For IBA models, all data are analyzed as one group;
%G1%
//Data set filename;
File = "..\Applied\longbeachP2.FF10.dat";
//Variable names;
Varnames = V1-V20;
//Sample size;
N = 328;

//Number of categories within each item (could specify different categories for different items);
Ncats(V1-V20) = 7;
//Number of categories within each item (could specify different categories for different items);
Model(V1-V20) = Graded(7);
//flexMIRT requires the item categories start with 0, so recoding is necessary if original
categories start with 1;
Code(V1-V20) = (1,2,3,4,5,6,7), (0,1,2,3,4,5,6);
//Total number of dimensions equals the number of primary factors plus the number of specific
factors;
Dimensions = 12;
//The number of primary factors (correlated);
Primary = 2;

<Constraints>
//All item slopes are initially set to zero;
Fix(V1-V20), Slope;
//Primary item slopes at Time-1 are freely estimated;
Free(V1-V10), Slope(1);
//Primary item slopes at Time-2 are freely estimated;
Free(V11-V20), Slope(2);

//Primary factor mean at Time-2, or Mean(2), is freely estimated;
Free G1, Mean(2);
//Correlation of primary factors at Time-1 and Time-2, or Cov(2,1), is freely estimated;
Free G1, Cov(2,1);
//Primary factor variance at Time-2, or Cov(2,2), is freely estimated;
Free G1, Cov(2,2);

//Primary parameters (slope and intercepts) of the anchor items are equated between time points;
Equal G1, (V1), Slope(1): G1, (V11), Slope(2);
Equal (V1, V11), Intercept(1);
Equal (V1, V11), Intercept(2);
```

```
Equal (V1, V11), Intercept(3);
Equal (V1, V11), Intercept(4);
Equal (V1, V11), Intercept(5);
Equal (V1, V11), Intercept(6);

//Specific item slopes associated with their corresponding specific factors are freely estimated;
Free(V1, V11), Slope(3);
Free(V2, V12), Slope(4);
Free(V3, V13), Slope(5);
Free(V4, V14), Slope(6);
Free(V5, V15), Slope(7);
Free(V6, V16), Slope(8);
Free(V7, V17), Slope(9);
Free(V8, V18), Slope(10);
Free(V9, V19), Slope(11);
Free(V10, V20), Slope(12);

//Each doublet of specific slopes was equated for identification purposes (unnecessary if more
than two time points are analyzed and specific factors are standardized);
Equal(V1, V11), Slope(3);
Equal(V2, V12), Slope(4);
Equal(V3, V13), Slope(5);
Equal(V4, V14), Slope(6);
Equal(V5, V15), Slope(7);
Equal(V6, V16), Slope(8);
Equal(V7, V17), Slope(9);
Equal(V8, V18), Slope(10);
Equal(V9, V19), Slope(11);
Equal(V10, V20), Slope(12);
```

Appendix D

The R function used for computing Wald statistics based on the IBA output from flexMIRT:

```
IBAWald <-

function(tp, cDef, nItems, nameCand, filePRM, fileCOV) {

  #tp = Number of time points. E.g., 3 is a proper input.
  #cDef = Contrast matrix defined. E.g., c(1, -1, 0, 1, 0, -1) is a proper input.
  #nItems = Number of items within each time point. E.g., 10 is a proper input.
  #nameCand = Names of the candidate items to be tested. E.g., c(2, 4, 6, 8) is a proper input.
  #filePRM = Filename of the parameter output from flexMIRT.
  #fileCOV = Filename of the covariance output from flexMIRT.

  ##################
  # Load 'Matrix' #
  ##################

  if("Matrix" %in% rownames(installed.packages()) == FALSE) {install.packages("Matrix")}
  require("Matrix")

  ###############
  # Omnibus DIF #
  ###############

  # v vectors
  temp <- read.table(filePRM,
                     sep = "\t",
                     nrows = tp*nItems,
                     fill = TRUE)[c(sapply(1:tp,
                                           function(x) nameCand + nItems*(x - 1))),
                                  -c(1:5)]
  cats <- temp[1:length(nameCand), 1]
  prms <- lapply(1:(tp*length(nameCand)),
                 function(x) head(as.numeric(temp[x, -1]), -nItems))
  prms <- lapply(split(mapply(function(x, y) c(x[1:y], sum(x[(y+1):(y+tp)])),
                              x = prms, y = rep(cats - 1, tp),
                              SIMPLIFY = FALSE),
                       rep(1:length(nameCand), tp)),
                 function(x) do.call(c, x))

  # C Matrices
  cMtx <- t(matrix(cDef, nrow = tp))
  cMtx.list <- mapply(function(x, y) sapply(x,
                                            function(z) diag(z, y),
                                            simplify = FALSE),
                      SIMPLIFY = FALSE,
                      x = rep(list(cMtx), length(nameCand)),
                      y = cats)

  cMtx.list <- lapply(lapply(cMtx.list,
                             function(x) lapply(split(x, rep(1:tp, each = nrow(cMtx))),
                                                function(y) do.call(rbind, y))),
                      function (z) do.call(cbind, z))

  # Sigma Matrices
  varcov <- read.csv(fileCOV, header = FALSE)
  varcov.order1 <- do.call(c,
                           mapply(FUN = c, SIMPLIFY = FALSE,
                                  mapply(FUN = function(x, y) seq.int(to = x,
                                                                      by = 1,
                                                                      length.out = y),
                                         x = cumsum(rep(cats, tp) - 1),
                                         y = (rep(cats, tp) - 1), SIMPLIFY = FALSE),
                                  seq.int(from = tail(cumsum(rep(cats, tp) - 1), 1) + 1,
                                          by = 1,
                                          length.out = tp*length(nameCand))))
  varcov <- as.matrix(varcov[ , -ncol(varcov)])[c(1:sum(rep(cats, tp))),
```

```
                                                c(1:sum(rep(cats, tp)))][varcov.order1,
                                                                         varcov.order1]

varcov.order2 <- mapply(FUN = function(x, y) seq.int(to = x,
                                                     by = 1,
                                                     length.out = y),
                        x = cumsum(rep(cats, tp)),
                        y = rep(cats, tp),
                        SIMPLIFY = FALSE)
sigma.list <- lapply(split(lapply(varcov.order2,
                                  FUN = function(x) varcov[x, x]),
                           c(rep(1:length(nameCand), tp))),
                     function(x) as.matrix(do.call(bdiag, x)))

# Calculate test statistics
Qstats <- mapply(FUN = function(c, v, sigma)
                       crossprod((c %*% v), solve(c %*% sigma %*% t(c))) %*% (c %*% v),
                 c = cMtx.list,
                 v = prms,
                 sigma = sigma.list)
df <- unlist(lapply(cMtx.list, nrow))
pval <- mapply(function(x, y) 1 - pchisq(x, y),
               x = Qstats,
               y = df)

Results <- list(data.frame("Q" = round(Qstats, 2),
                           "df" = df, "p" = round(pval, 3),
                           row.names = paste0("Item", nameCand)))
names(Results) <- "Omnibus"

###############
# By Contrast #
###############

if (nrow(cMtx) > 1) {
  bycontrast <- list()
  for (i in 1:nrow(cMtx)) {
    c.pw <- as.numeric(cMtx[i, ]) # contrast row
    ind.pw <- as.numeric(which(c.pw != 0))

    # v Vectors
    label.pw <- lapply(split(mapply(function(x, y) seq.int(to = x*y,
                                                           by = 1,
                                                           length.out = y),
                                    SIMPLIFY = FALSE,
                                    x = ind.pw,
                                    y = rep(cats, each = length(ind.pw))),
                             rep(1:length(nameCand), each = length(ind.pw))),
                       function(z) do.call(c, z))
    prms.pw <- mapply(function(x, y) unlist(x)[y],
                      x = prms,
                      y = label.pw,
                      SIMPLIFY = FALSE)

    # C Matrices
    cMtx.list.pw <- lapply(mapply(function(x, y) sapply(x,
                                                        function(z) diag(z, y),
                                                        simplify = FALSE),
                                  SIMPLIFY = FALSE,
                                  x = rep(list(c.pw[ind.pw]), length(nameCand)),
                                  y = cats),
                           function (x) do.call(cbind, x))

    # Sigma Matrices
    varcov.order.pw1 <- do.call(c,
                                lapply(which(c.pw != 0),
                                       function(x) seq.int(to = tail(cumsum(cats), 1)*x,
                                                           by = 1,
                                                           length.out = tail(cumsum(cats), 1))))
    varcov.pw <- varcov[varcov.order.pw1, varcov.order.pw1]
    varcov.order.pw2 <- mapply(FUN = function(x, y) seq.int(to = x,
```

```
                                                          by = 1,
                                                          length.out = y),
                                SIMPLIFY = FALSE,
                                x = cumsum(rep(cats, length(ind.pw))),
                                y = rep(cats, length(ind.pw)))
      sigma.list.pw <- lapply(split(lapply(varcov.order.pw2,
                                           FUN = function(x) varcov.pw[x, x]),
                                    c(rep(1:length(nameCand), length(ind.pw)))),
                              function(x) as.matrix(do.call(bdiag, x)))

      # Calculate test statistics
      Qstats.pw <- mapply(FUN = function(c, v, sigma)
                            crossprod((c %*% v), solve(c %*% sigma %*% t(c))) %*% (c %*% v),
                          c = cMtx.list.pw,
                          v = prms.pw,
                          sigma = sigma.list.pw)
      df.pw <- unlist(lapply(cMtx.list.pw, nrow))
      pval.pw <- mapply(function(x, y) 1 - pchisq(x, y),
                        x = Qstats.pw,
                        y = df.pw)

      bycontrast[[i]] <- data.frame("Q" = round(Qstats.pw, 2),
                                    "df" = df.pw,
                                    "p" = round(pval.pw, 3),
                                    row.names = paste0("Item", nameCand))
    } #End of 'for' loop
    Results <- c(Results, bycontrast)
    names(Results) <- c("Omnibus", paste0("Contrast ", 1:nrow(cMtx)))
  } #End of 'By Contrast'
  return(Results)
} #End of function IBAWald()
```

To use the above IBAWald function in R for the 10-item MFQ subscale across two waves:

```
source("IBAWald.R")

IBAWald(tp = 2, cDef = c(1, -1), nItems = 10, nameCand = c(2:10),
        filePRM = "./Applied/longbeachP2.FF10-prm.txt",
        fileCOV = "./Applied/longbeachP2.FF10-cov.txt")
```

To use the above IBAWald function in R for the simulated 10-item test across three waves:

```
source("IBAWald.R")

IBAWald(tp = 3, cDef = c(1, -1, 0,
                         1, 0, -1), nItems = 10, nameCand = c(3:10),
        filePRM = "./Simulated/Replication0001-prm.txt",
        fileCOV = "./Simulated/Replication0001-cov.txt")
```

Bibliography

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*(4), 443-459. doi: 10.1007/BF02294168

Bowers, E. P., Li, Y., Kiely, M. K., Brittian, A., Lerner, J. V., Lerner, R. M. (2010). The Five Cs model of positive youth development: A longitudinal analysis of confirmatory factor structure and measurement invariance. *Journal of Youth and Adolescence, 39*(7), 720-735. doi: 10.1007/s10964-010-9530-9

Cai, L. (2008). SEM of another flavour: Two new applications of the supplemented EM algorithm. *British Journal of Mathematical and Statistical Psychology, 61(2),* 309-329. doi: 10.1348/000711007X249603

Cai, L. (2010). A two-tier full-information item factor analysis model with applications. *Psychometrika, 75*(4), 581-612. doi: 10.1007/s11336-010-9178-0

Cai, L. (2015). flexMIRT® version 3: Flexible multilevel and multidimensional item analysis and test scoring [Computer software]. Chapel Hill, NC: Vector Psychometric Group, LLC.

Cai, L., Yang, J. S., & Hansen, M. (2011). Generalized full-information item bifactor analysis. *Psychological Methods, 16*(3), 221-248. doi: 10.1037/a0023350

Coertjens, L., Donche, V., De Maeyer, S., Vanthournout, G., & Van Petegem, P. (2012). Longitudinal measurement invariance of Liert-type learning strategy scales: Are we using the same ruler at each wave? *Journal of Psychoeducational Assessment, 30*(6), 577-587.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B*

(Methodological), 39(1), 1-38. Retrieved from http://www.jstor.org.www2.lib.ku.edu/stable/2984875

Dobson, A. J., & Barnett, A. (2008). *An introduction to generalized linear models (3rd ed.).* New York, NY: Chapman and Hall/CRC.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, N.J.

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika, 57*(3), 423-436. Retrieved from http://search.proquest.com/docview/618256604?accountid=14556

Hill, C. D. (2006). *Two models for longitudinal item response data* (Doctoral dissertation). Available from PsycINFO. (621568031; 2006-99016-148). Retrieved from http://search.proquest.com/docview/621568031?accountid=14556

Holzinger, K. J., & Swineford, F. (1937). The bi-factor method. *Psychometrika, 2*(1), 41-54.

Kim, S.-H., & Cohen, A. S. A comparison of linking and concurrent calibration under the graded response model. *Applied Psychological Measurement, 26*(1), 25-41. doi: 10.1177/0146621602026001002

Kim, S.-H., Cohen, A. S., & Park, T. (1995). Detection of differential item functioning in multiple groups. *Journal of Educational Measurement, 32*(3), 261-276. Retrieved from http://www.jstor.org.www2.lib.ku.edu/stable/1435297

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices (3rd ed.).* New York, NY: Springer Science + Business Media. doi: 10.1007/978-1-4939-0317-7

Langer, M. M. (2008). *A reexamination of Lord's Wald test for differential item functioning using item response theory and modern error estimation* (Doctoral dissertation). Retrieved from

http://search.proquest.com/docview/622062303?accountid=14556. (622062303; 2009-99100-046).

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 22*(140), 5-55.

Little, T. D. (2013). *Longitudinal structural equation modeling*. New York, NY: The Guilford Press.

Lord, F. M. (1977). A study of item bias using item characteristic curve theory. In Y. H. Poortinga, *Basic problems in cross-cultural psychology* (pp. 19-29). Amsterdam: Swets and Zeitlinger.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum.

Meade, A. W., & Wright, N. A. (2012). Solving the measurement invariance anchor item problem in item response theory. *Journal of Applied Psychology, 97*(5), 1016-1031. doi: 10.1037/a0027934

Meng, X.-L., & Rubin, D. B. (1991). Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. *Journal of the American Statistical Association, 86(416),* 899-909. doi:10.1080/01621459.1991.10475130

Mukherjee, S., Gibbons, L. E., Kristjansson, E., & Crane, P. K. (2013). Extension of an iterative hybrid ordinal logistic regression/item response theory approach to detect and account for differential item functioning in longitudinal data. *Psychological Test and Assessment Modeling, 55*(2), 127-147.

Paek, I., & Cai, L. (2014). A comparison of item parameter standard error estimation procedures for unidimensional and multidimensional item response theory modeling. *Educational and Psychological Measurement, 74*(1), 58-76. doi: 10.1177/0013164413500277

Pitts, S. C., West, S. G., & Tein, J.-Y. (1996). Longitudinal measurement models in evaluation research: Examining stability and change. *Evaluation and Program Planning, 19*(4), 333-350. doi: 10.1016/S0149-7189(96)00027-4

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Rosenberg, M. (1965). *Society and the adolescent self-image.* Princeton, N.J.: Princeton University Press.

Samejima, F. (1970). Estimation of latent ability using a response pattern of graded scores. *Psychometrika, 35*(1), 139-139. doi: 10.1007/BF02290599

Short, S. D. (2014). *Power of alternative fit indices for multiple group longitudinal tests of measurement invariance* (Doctoral dissertation). Retrieved from https://kuscholarworks.ku.edu/handle/1808/14570

Tian, W., Cai, L., Thissen, D., & Xin, T. (2012). Numerical differentiation methods for computing error covariance matrices in item response theory modeling: An evaluation and a new proposal. *Educational and Psychological Measurement, 73*(3), 412-439. doi: 10.1016/S0149-7189(96)00027-4

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society, 54*(3), 426-482. doi: 10.1090/S0002-9947-1943-0012401-3

Wang, W.-C., Shih, C.-L., Sun, G.-W. (2012). The DIF-free-then-DIF strategy for the assessment of differential item functioning. *Educational and Psychological Measurement, 72*(4), 687-708. doi: 10.1177/0013164411426157

Wang, W., & Yeh, Y. (2003). Effects of anchor item methods on differential item functioning detection with the likelihood ratio test. *Applied Psychological Measurement, 27*(6), 479-498. doi: 10.1177/0146621603259902

Woods, C. M. (2009). Empirical selection of anchors for tests of differential item functioning. *Applied Psychological Measurement, 33*(1), 42-57. doi: 10.1177/0146621607314044

Woods, C. M. (2011). DIF testing for ordinal items with poly-SIBTEST, the Mantel and GMH tests, and IRT-LR-DIF when the latent distribution is nonnormal for both groups. *Applied Psychological Measurement, 35*(2), 145-164. doi: 10.1177/0146621610377450

Woods, C. M., Cai, L., & Wang, M. (2013). The Langer-improved Wald test for DIF testing with multiple groups: Evaluation and comparison to two-group IRT. *Educational and Psychological Measurement, 73*(3), 532-547. doi: 10.1177/0013164412464875

Wu, C.-H., Chen, L. H., & Tsai, Y.-M. (2009). Longitudinal invariance analysis of the satisfaction with life scale. *Personality and Individual Differences, 46*(4), 396-401. doi: 10.1016/j.paid.2008.11.002

Zelinski, E. M., & Gilewski, M. J. (2004). A 10-item Rasch modeled memory self-efficacy scale. *Aging & Mental Health, 8*(4), 293-306. doi: 10.1080/13607860410001709665

Zelinski, E. M., Gilewski, M. J., & Anthony-Bergstone, C. R. (1990). Memory functioning questionnaire: Concurrent validity with memory performance and self-reported memory failures. *Psychology and Aging, 5*(3), 388-399. doi: 10.1037/0882-7974.5.3.388

Zelinski, E. M., & Kennison, R. F. (2011). Long Beach Longitudinal Study . ICPSR26561-v2. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2011-06-17. http://doi.org/10.3886/ICPSR26561.v2